



**UNIVERSIDAD NACIONAL AUTÓNOMA DE
MÉXICO**

FACULTAD DE QUÍMICA

**PROGRAMA DE MAESTRÍA Y DOCTORADO EN
CIENCIAS BIOQUÍMICAS**

DEPARTAMENTO DE BIOQUÍMICA

**Evolución molecular de la familia de factores de inicio
de la traducción eIF4E**

T E S I S

Que para obtener el grado de:

MAESTRO EN CIENCIAS (BIOQUÍMICAS)

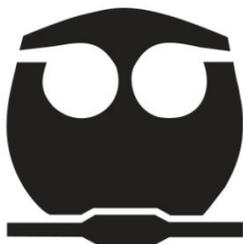
P R E S E N T A

Biól. Miguel Ángel Cruz Espindola

Director de tesis:

Dr. León Patricio Martínez Castilla

México, D.F. 2011



UNAM
POSGRADO 
Ciencias Bioquímicas



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Este trabajo se realizó bajo la dirección de Dr. León Patricio Martínez Castilla en el Departamento de Bioquímica (Conjunto E) de la Facultad de Química de la Universidad Nacional Autónoma de México, y se desarrolló con el apoyo de los proyectos PAPIIT-DGAPA: IN21160 y PAIP-6290-07 (Facultad de Química). El alumno Miguel Ángel Cruz Espindola contó con el apoyo económico de CONACYT, mediante la beca 215188.

El comité tutorial que asesoró el desarrollo de este proyecto estuvo integrado por:

Dr. León Patricio Martínez Castilla	Facultad de Química, UNAM.
Dra. Tzvetanka Dimitrova Dinkova	Facultad de Química, UNAM.
Dr. Lorenzo Patrick Segovia Forcella	Instituto de Biotecnología, UNAM

Agradecimientos

A mi familia por su comprensión y apoyo incondicional.

Al Dr. León Patricio Martínez Castilla por permitirme realizar en su laboratorio mi maestría y guiarme durante el desarrollo del proyecto.

A la Dra. Tzvetanka y al Dr. Lorenzo por las críticas y aportaciones que enriquecieron este trabajo.

A los miembros de mi jurado de examen: Dra. Tzvetanka Dimitrova, Dr. Miguel Ángel Cevallos, Dr. Ricardo Alfredo Grande, Dr. Arturo Carlos II Becerra y el Dr. Víctor Manuel Valdés, por las valiosas correcciones.

A mis compañeros del laboratorio: Alejandra y Quetzal por las sugerencias realizadas a este trabajo y la ayuda que recibí de ellos.

A mis amigos: Marisela, Jonathan y Gabriel por su amistad, consejos, las risas y los momentos agradables que hemos compartido.

Al Dr. Edgar Vázquez por la ayuda y consejos.

Al CONACYT por el apoyo económico.

A mi universidad, la UNAM, por la formación recibida en ella.

Contenido

1 Resumen	5
2 Introducción	6
2.1 El Factor de Inicio de la Traducción eIF4E	7
2.2 Filogenias Moleculares	9
2.3 Detección de Selección Natural a Nivel Molecular	20
2.4 Antecedentes	24
3 Hipótesis y Objetivo	29
4 Material y Métodos	30
4.1 Reconstrucción filogenética de la familia de factores de inicio de la traducción eIF4E	31
4.2 Enraizamiento del árbol filogenético	34
4.3 Detección de Sitios Bajo Selección Natural	37
5 Resultados	38
5.1 Reconstrucción filogenética de la familia de factores de inicio de la traducción eIF4E	38
5.2 Enraizamiento del árbol filogenético	54
5.3 Detección de Sitios Bajo Selección Natural	79
5.4 Detección de Selección Natural en eIF(iso)4E de Viridiplantae	84
5.5 Detección de Selección Natural en nCBP de Viridiplantae	89
6 Discusión	91
6.1 Reconstrucción filogenética de la familia de factores de inicio de la traducción eIF4E	91
6.2 Enraizamiento del árbol filogenético	96
6.3 Detección de Sitios Bajo Selección Natural	101
6.4 Detección de Selección Natural en eIF(iso)4E de Viridiplantae	103
6.5 Detección de Selección Natural en nCBP de Viridiplantae	109
7 Conclusiones	111
8 Perspectivas	113
9 Bibliografía	115

1. Resumen

La traducción es uno de los principales procesos celulares, mediante el cual se producen todas las proteínas; la traducción, en los 3 dominios celulares presenta un paso característico (el reclutamiento de la subunidad menor ribosomal al RNAm), sin embargo el cómo se lleva a cabo este paso en cada uno de los dominios ha variado. En los Eucariontes, el inicio de la traducción requiere de una gran cantidad de proteínas, entre las cuales sobresale el factor de inicio de la traducción 4E (eIF4E) por su relevancia funcional, ya que reconoce y se une a la estructura *cap*. Inicialmente se creía que la familia de eIF4E estaba conformada por una sola proteína, sin embargo, con el descubrimiento de otros miembros de esta familia, esta visión cambió. Algunos análisis filogenéticos permitieron clasificar a los distintos miembros de esta familia dentro de 3 clases estructurales, mediante los residuos de aminoácidos que se presentaban en las posiciones 43 y 56, de acuerdo a la numeración de eIF4E de *Homo sapiens*. La clase estructural I incluye a las secuencias que presentan en las posiciones 43 y 56 un residuo de triptófano y que participan de manera activa en el inicio de la traducción; la clase estructural II consiste de secuencias que presentan en la posición 43 residuos de leucina, fenilalanina o tirosina y en la posición 56 tirosina o fenilalanina, y se ha observado que estas proteínas se unen a la estructura *cap*, pero algunas no pueden interactuar con eIF4G y participar en la traducción; la clase estructural III agrupa a las secuencias que presentan residuos de cisteína o tirosina en la posición 56, mientras que en la posición 43 presentan residuos de triptófano, estas proteínas pueden interactuar con eIF4G pero débilmente con la estructura *cap*. Las diferencias en los patrones de expresión aunadas a las diferentes afinidades de unión, permiten establecer diferencias funcionales entre los miembros de eIF4E respecto a la función tradicional y característica de esta familia, el reconocimiento y la unión a *cap*. La inclusión de las secuencias de más miembros de esta familia, y el empleo de métodos de reconstrucción filogenética más robustos, permitieron la obtención de un árbol filogenético con una topología distinta al obtenido a un análisis anterior (Joshi et al 2005), en el cual se observa la formación de 2 grandes clados, el primero incluye a las secuencias pertenecientes a la clase estructural I, mientras que en el segundo clado se encuentran las secuencias de la clase estructural II y III; así mismo detectamos secuencias pertenecientes a la clase estructural III del linaje Fungi, y la ausencia de secuencias de la clase estructural II en el mismo linaje, anteriormente no reportado. Por otra parte las pruebas de detección de selección natural positiva fueron negativas en el origen de cada una de las clases, permitiendo plantear que el mantenimiento de los miembros duplicados de eIF4E se deba a una neo-funcionalización mediante el mecanismo evolutivo de la deriva génica u otro mecanismo neutral y no a la selección natural positiva o adaptativa. Como un caso particular, los miembros de la clase estructural I del linaje de Viridiplantae, se han mantenido posiblemente mediante un proceso de sub-funcionalización; por otra parte los miembros de la clase estructural II, del linaje de Viridiplantae se han mantenido por el mecanismo de neo-funcionalización posiblemente por algún proceso neutral debido a la ausencia de sitios bajo selección natural. Las reconstrucciones filogenéticas y las pruebas de detección de selección, nos permiten obtener una imagen más completa sobre la evolución de esta familia de proteínas.

2. Introducción

La traducción es uno de los principales procesos celulares, mediante el cual se lleva a cabo la síntesis de proteínas, y es fundamental para el crecimiento, el mantenimiento y la diferenciación celular, y forma parte del proceso de expresión génica que consiste en la decodificación de la información contenida en el DNA en moléculas: proteínas o RNAs, necesarias para la célula.

Todos los transcritos de la RNA polimerasa II son modificados cotranscripcionalmente mediante la adición de una 7-metilguanosina en el extremo 5', a este grupo se le conoce como el "cap". Esta modificación juega un papel importante en el *splicing*, la poliadenilación y la exportación nuclear (McKendrick *et al.*, 1999). La presencia de un *cap* en el extremo 5' del RNAm estimula la traducción, además de estabilizar el RNAm contra la degradación (Jankowska-Anyszka *et al.*, 1998). El primer paso crítico durante el inicio de la traducción es el reclutamiento del complejo de pre-iniciación, al extremo terminal 5'RNAm (Ptushkina *et al.*, 2004 y Von der Haar *et al.*, 2000). Este reclutamiento es un proceso de varios pasos que involucra al menos 10 diferentes factores de iniciación de la traducción eIFs. La unión del complejo de pre-iniciación 43S (subunidad ribosomal 40S cargada con el RNAt iniciador metionina) al extremo 5' del RNAm, lo cataliza un complejo de factores de iniciación, eIF4F. El eIF4F es un complejo ternario, que consiste de eIF4E (Combe *et al.*, 2005), la subunidad de reconocimiento y unión de la estructura 5' *cap* 7-metilguanosina (Joshi *et al.*, 2004), eIF4A, una RNA helicasa con caja DEAD, y eIF4G, un polipéptido que interactúa con eIF4A, eIF4E y con la subunidad ribosomal 40S vía contactos con eIF3 (Combe *et al.*, 2005), (figura 1).

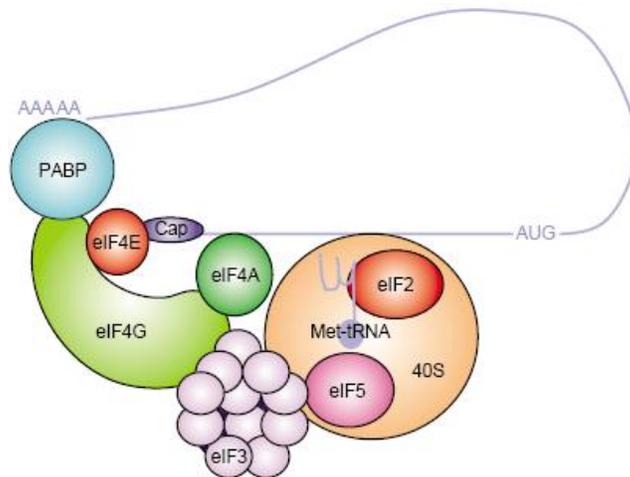


Figura 1. Complejo eucariótico del inicio de la traducción. (Robaglia y Caranta, 2006)

eIF4F se une al 5' *cap* del RNAm y desenreda la estructura secundaria dentro de la región no traducible (UTR) 5' para facilitar el reclutamiento y el escaneo del complejo de pre-iniciación 43S. También se piensa que eIF4F puede promover la reiniciación de la traducción, mediante la circularización del RNAm a través de interacciones entre eIF4G y las proteínas de unión a poliA, PABP. En adición a eIF4F, en las plantas se ha observado la presencia de un segundo complejo de unión a “*cap*” llamado eIFiso4F, el cual está compuesto de diferentes subunidades eIF4E y eIF4G. eIFiso4F consiste de eIFiso4G y una subunidad de unión a “*cap*” llamado eIFiso4E. La actividad de eIF4F es regulada mediante la subunidad de unión al *cap*, eIF4E (Combe *et al.*, 2005).

2.1 El Factor de Inicio de la Traducción eIF4E

Como se mencionó anteriormente el factor eIF4E interactúa con el RNAm y con el factor de iniciación eIF4G; la interacción de eIF4E con la estructura *cap* potencia la afinidad de eIF4E por eIF4G y viceversa (Joshi *et al.*, 2004). El componente mayor de eIF4F, eIF4G sirve como un adaptador multipropósito capaz de reclutar un número de actividades necesarias hacia el extremo 5' de el RNAm (Von der Haar *et al.*, 2000) ya que tiene sitios de unión no sólo para eIF4E, sino también para eIF3, eIF4A y la proteína de unión al poliA PABP (Ptushkina *et al.*, 2004). La unión a la RNA helicasa eIF4A es necesaria, para romper la estructura secundaria en la región 5' no traducible (Von der Haar *et al.*, 2000).

La estructura de los cristales de eIF4E de *Mus musculus*, *Homo sapiens* y la estructura resuelta de eIF4E de *Saccharomyces cerevisiae*, en cada uno de los casos unido a análogos de “cap”, muestran que el factor eIF4E consiste de 8 láminas β y de 3 α -hélices formando la palma y el dorso de una mano en forma de copa respectivamente. El cap 7-metilguanosa permanece en un bolsillo hidrofóbico sobre la cara cóncava (McKendrick *et al.*, 1999), el reconocimiento del cap se lleva a cabo vía la interacción de dos residuos de triptófano altamente conservados (posición 56 y 102 en eIF4E de *Homo sapiens*), (figura 2). Un tercer residuo de Trp (Trp166 en eIF4E de humano) reconoce la presencia del grupo N7-metil de la estructura cap (Joshi *et al.*, 2004).

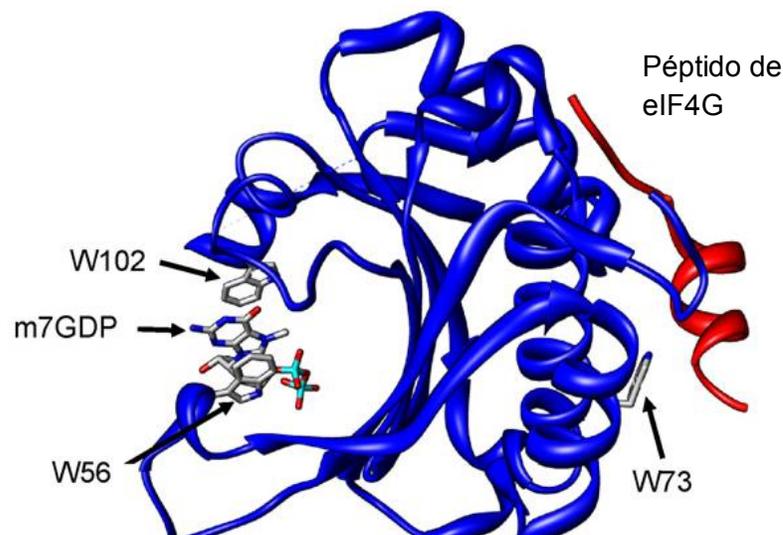


Figura 2. Estructura tridimensional de eIF4E de ratón, en complejo con m7GDP y un péptido de eIF4G (PDB 1EJ4). Los residuos importantes para la unión del cap (W56 and W102) y el ligando (W73) se muestran en representación de esqueleto (Goodfellow y Roberts, 2008)

La actividad de eIF4E es regulada mediante proteínas de unión a eIF4E. En mamíferos estas proteínas son conocidas como 4E-BP1, 4E-BP2, y 4E-BP3. Una proteína con propiedades similares a 4E-BP se encuentra en *Saccharomyces cerevisiae*, la llamada p20. Las 4E-BPs interactúan con eIF4E para inhibir la asociación de eIF4E con eIF4G (Joshi *et al.*, 2004). 4E-BP hipofosforilado secuestra a eIF4E, inhibiendo la traducción, cuando 4E-BP está hiperfosforilado no se une a eIF4E y éste queda libre para participar en el inicio de la traducción. El 4E-BP compite con eIF4G para unirse a eIF4E, debido a que ellos comparten el mismo sitio de unión sobre eIF4E (Goodfellow y Roberts, 2008).

Las 4E-BPs y eIF4Gs contienen una región consenso de unión a eIF4E, TyrXXXXLeuΦ, donde Φ es un aminoácido hidrofóbico y X es cualquier aminoácido. Los estudios cristalográficos de eIF4E han mostrado que la His-37, Pro-38, Val-69, Trp-73, Leu-131, Glu-132 y Leu-35 (numeración basada en eIF4E de humano) interactúan con la región de unión a eIF4E dentro eIF4G y 4E-BPs. Los residuos de aminoácidos Val-69 y Trp-73 se encuentran dentro de una secuencia consenso conservada filogenéticamente (Ser/Thr)Val(Glu/Asp)(Glu/Asp)PheTrp, donde los residuos ácidos no están completamente conservados. La sustitución del Trp-73 de eIF4E en mamíferos por un aminoácido no aromático rompe la capacidad de eIF4E para interactuar tanto con eIF4G como con 4E-BPs. La sustitución de Val-69 por un residuo de Gly resulta en la producción de una variante de eIF4E que se une a 4E-BP1 eficientemente pero compromete su capacidad para unirse tanto a eIF4G como a 4E-BP2 (Joshi *et al.*, 2004).

2.2 Filogenias Moleculares

La biología evolutiva está basada en el concepto de que los organismos comparten un origen común y que subsecuentemente divergieron en el tiempo. Las filogenias representan nuestro intento de reconstruir su historia evolutiva, y probablemente ha tomado más interés hoy en día que en el pasado. Por años, la filogenética tenía un papel menos importante en la biología evolutiva y, no hace mucho, la importancia de las filogenias se ha reconocido en la mayoría de los campos de la biología (Huelsenbeck y Rannala, 1997).

La filogenética molecular al igual que la genómica comparada emplean alineamientos múltiples de secuencia moleculares, de proteínas o de ácidos nucleicos, para estudiar cómo han cambiado las secuencias a través del tiempo, convirtiendo la información contenida dentro de las secuencias en arboles evolutivos o filogenéticos (Page y Holmes, 1998); el principal objetivo de algunos estudios, es el encontrar el árbol filogenético que mejor describa la relación evolutiva de las secuencias en estudio. Por otra parte, los árboles también han proporcionado conocimiento valioso en estudios moleculares y fisiológicos, incluyendo cómo el repertorio de proteínas evolucionó; la evolución del genoma; y el desarrollo de clasificaciones taxonómicas y conceptos de especie. Una reconstrucción exacta de las relaciones evolutivas es importante también, no sólo en los casos anteriores, sino en aquellos estudios donde el principal objetivo no es la inferencia

del árbol, como cuando se investiga la presión de selección sobre proteínas o la identificación de elementos conservados en el genoma (Keith, 2008).

Una filogenia o árbol filogenético es una representación gráfica de las relaciones genealógicas entre especies, entre genes, entre poblaciones, o entre individuos. Los matemáticos definen un grafo como un conjunto de vértices y un conjunto de aristas que los conectan, y un árbol estaría definido como un grafo conectado sin asas. En filogenética a menudo se usan las palabras “nodos” para referirse a los vértices y “ramas” para las aristas. Las puntas, hojas o nodos externos representan las especies, secuencias, genes actuales, mientras que los nodos internos usualmente representan ancestros extintos para los cuales no hay secuencias disponibles. El ancestro para todas las secuencias es la raíz del árbol (Yang, 2006); el conocer la ubicación de la raíz del árbol, nos permite determinar el orden de los eventos de divergencia (Keith, 2008).

Existen actualmente distintos métodos para la reconstrucción de las filogenias los cuales pueden clasificarse de diversas maneras de acuerdo a sus características. Algunos métodos están basados en distancia y en dichos métodos la distancia se calcula mediante la comparación pareada de secuencias y la generación de una matriz de distancia, la cual se usa en análisis subsecuentes. A menudo se usa un algoritmo de agrupamiento para convertir la matriz de distancia en un árbol filogenético. Los métodos más populares son *Unweighted Pair Group Method using Arithmetic Averages* (UPGMA) y el *neighbour-joining* (NJ). Otros métodos se basan en caracteres e intentan ajustar los caracteres (nucleótidos o aminoácidos) observados en todas las especies y en cada uno de los sitios a un árbol. Los métodos de máxima parsimonia, de máxima verosimilitud y los métodos bayesianos se basan en caracteres. Los métodos de distancia son a menudo computacionalmente más rápidos que los métodos basados en caracteres, y pueden emplearse para analizar diferentes tipos de datos. Los métodos también pueden separarse en los que construyen un árbol siguiendo un algoritmo (métodos de agrupación) o en los que se basan por el criterio de optimización que usan (métodos de búsqueda). El UPGMA y el *neighbour-joining* pueden clasificarse dentro de los primeros ya que usan un algoritmo de agrupación para llegar a un solo árbol desde los datos. Los métodos basados en la optimización usan un determinado criterio de optimización para medir el ajuste del árbol a los datos, y el árbol con la puntuación óptima se estima como el árbol real (Yang, 2006).

Las aproximaciones más comunes para la reconstrucción filogenéticas son el algoritmo *neighbour-joining* (NJ) y la búsqueda de árboles que usan un criterio de optimización como lo es la máxima parsimonia (MP), la máxima verosimilitud (ML) y los métodos bayesianos (Holder y Lewis, 2003).

Neighbour-joining es un algoritmo extremadamente popular y es relativamente rápido, y como todos los métodos, tienen un buen desempeño cuando la divergencia entre las secuencias es baja. El primer paso en el algoritmo es convertir las secuencias de DNA o de proteínas en una matriz de distancia que representa un estimado de la distancia evolutiva entre las secuencias (la distancia evolutiva es el número de cambios que han ocurrido a lo largo de la rama, entre 2 secuencias). Una potencial y debilidad seria para los métodos de distancia como *neighbour-joining*, es que las diferencias observadas entre secuencias no sean precisamente reflejo de la distancia evolutiva entre ellas. Sustituciones múltiples en el mismo sitio oscurecen la distancia verdadera y hace que las secuencias parezcan artificialmente cercanas una a otra (Holder y Lewis, 2003). Posteriormente se construye un árbol inicial en forma de estrella, a continuación se agrupan los 2 nodos terminales con la menor distancia para forma un nodo interno, se calcula la longitud de la rama del nodo interno hacia los 2 nodos externos; se crea una nueva matriz de distancia, en la cual se calcula la distancia del nodo interno hacia los demás nodos terminales, exceptuando aquellos nodos que agrupa el nodo interno, y se repite el procedimiento de agrupación de nodos terminales con la menor distancia hasta que todos los nodos terminales estén agrupados por nodos internos (Saitou y Nei, 1987). *Neighbour-joining* es un estimador extremadamente rápido de árboles filogenéticos, lo que lo hace relativamente interesante cuando se emplean datos claros (por ejemplo, secuencias que han divergido recientemente). Cuando el objetivo es inferir relaciones antiguas, puede ser difícil llegar a los valores reales para la matriz de distancia que es la entrada para el *neighbour-joining*. Obviamente si los datos de entrada para un algoritmo son erróneos, tiene poco oportunidad de obtener éxito (Holder y Lewis, 2003).

Los métodos de optimización o de búsqueda se basan sobre el principio de optimización o de criterio óptimo. Bajo este principio, el árbol preferido se determina mediante la asignación de una puntuación de optimización a todas las posibles topologías (o todas las

topologías potencialmente correctas) de acuerdo a cierto procedimiento y eligiendo la topología que muestre la puntuación óptima alta o baja (Tabla1) (Yang, 2006)

Tabla 1. Criterio de optimización usado para reconstrucción de filogenias (Yang, 2006)

Método	Criterio
Máxima Parsimonia	Número mínimo de cambios, minimizados sobre el estado ancestral
Máxima Verosimilitud	Puntuación del logaritmo de la verosimilitud, optimizada sobre la longitud de rama y los parámetros del modelo
Mínima Evolución	Longitud del árbol (Suma de la longitud de ramas, a menudo estimada mediante mínimos cuadrados)
Bayesiano	Probabilidad posterior, calculada mediante la integración de las longitudes de ramas y parámetros de sustitución

Dentro de los métodos que emplean el criterio de optimización encontramos el método de máxima parsimonia. El método de máxima parsimonia se desarrolló originalmente para caracteres morfológicos (Nei y Kumar, 2000); sin embargo, también puede emplearse para datos moleculares, si cada una de las columnas de un alineamiento de secuencias es tratado como un carácter (Higgs y Attwood, 2005). La idea central en la que se basa y se desarrolló el método de máxima parsimonia es que “la mejor hipótesis para explicar un proceso es aquella que requiere el menor número de suposiciones” (Nei y Kumar, 2000); por lo tanto, el criterio que emplea el método de máxima parsimonia para la elección de la topología correcta del árbol filogenético, para un alineamiento de secuencias, es aquella que presente el menor número de cambios de carácter (Salemi y Vandamme, 2003).

Algunas ventajas del método de máxima parsimonia es que aparentemente hace pocas suposiciones acerca de los procesos evolutivos, es relativamente fácil de entender, y que matemáticamente se ha estudiado extensamente; sin embargo, el criterio de elección del árbol más parsimonioso es sujeto a considerables controversias (Page y Holmes, 1998). Como principal desventaja es que son sensibles al fenómeno de atracción de ramas largas o de atracción de ramas cortas, en el cual las ramas largas o las cortas en la

topología correcta tienden a unirse o atraerse en el árbol reconstruido, y originan una topología incorrecta (Nei y Kumar, 2000).

Otro método que emplea el criterio de optimización es la máxima verosimilitud. El método de máxima verosimilitud se acredita usualmente al estadístico inglés R. A. Fisher, quien describió el método en 1932 y fue el primero en investigar sus propiedades. El método de máxima verosimilitud depende de la especificación completa de los datos y de un modelo de probabilidad para describir los datos. La probabilidad de observar los datos bajo el modelo asumido, cambia dependiendo de los valores de los parámetros del modelo. El método de máxima verosimilitud escoge el valor de los parámetros que maximicen la probabilidad de observar los datos. La aplicación de la estimación de la máxima verosimilitud para el problema de filogenias la sugirió inicialmente Edwards y Cavalli-Sforza, en los inicios de 1960. Sin embargo, ellos percibieron la dificultad computacional en el momento e intentaron métodos de aproximación en su lugar. La subsecuente historia de la máxima verosimilitud puede leerse como un constante progreso en el cual las barreras computacionales se rompieron y los modelos usados se hicieron biológicamente más realistas (Huelsenbeck y Crandall, 1997).

En contraste con el método de máxima parsimonia, los métodos de máxima verosimilitud usan el logaritmo natural del valor de la verosimilitud ($\ln L$) para cada una de las topologías como el criterio de puntuación que se busca a optimizar: el árbol que muestre el $\ln L$ más alto es el que se elige como el árbol óptimo (Takahashi y Nei, 2000). La verosimilitud es la probabilidad de los datos dado el modelo; y por lo tanto en evolución molecular los datos son el alineamiento de secuencias, y el modelo, en un sentido amplio, es el árbol que relaciona las secuencias más el mecanismo de cambio molecular. Es usual separar las dos partes del modelo, y llamar a la parte del árbol “el árbol” y la parte del mecanismo “el modelo”. La parte del mecanismo, el modelo, es una idea de la forma en que las secuencias moleculares cambian sobre el tiempo (Foster, 2001). En máxima verosimilitud una hipótesis es juzgada de acuerdo a que tan bien predice los datos observados; el árbol que tiene una alta probabilidad de representar las secuencias observadas es preferido (su probabilidad es la verosimilitud del árbol). Para usar este método, debemos ser capaces de calcular la probabilidad de un conjunto de datos dado un árbol filogenético (Holder y

Lewis, 2003). A menudo se ha considerado que una de las ventajas de la máxima verosimilitud sobre los métodos basados en parsimonia es que permite el uso de modelos explícitos de evolución apropiados para el conjunto de datos que se están examinando (Keane *et al.*, 2006). Los modelos de evolución de proteínas o de reemplazamiento de aminoácidos, describen las probabilidades de cambio de un aminoácido por otro y es una herramienta indispensable para la caracterización del proceso de evolución de proteínas. Estos modelos proporcionan la base para la reconstrucción de filogenias de proteínas bajo la máxima verosimilitud y los métodos bayesianos. Dayhoff y colaboradores introdujeron el modelo de reemplazamiento de aminoácidos más influyente, un modelo de Markov homogéneo reversible en el tiempo de 20 estados. Una matriz de reemplazamiento de 20 estados presenta un gran número de parámetros y la estimación de esos parámetros anteriormente se obtenía a partir de un gran conjunto de bases de datos. Debido a esto, se han propuesto diferentes matrices empíricas con tasas fijas relativas de reemplazamiento de aminoácidos, como la matriz de Dayhoff (Dayhoff *et al.*, 1978), o las matrices JTT (Jones *et al.*, 1992), mtREV (Adachi y Hasegawa, 1996) o WAG (Whelan y Goldman, 2001). Estos modelos generalmente asumen que el proceso de reemplazamiento de aminoácidos es muy similar a través de todas las posiciones, y consideran que la función y la estructura de la proteína imponen límites sobre cuál posición puede cambiar. Esta información evolutiva puede inferirse bajo la consideración de que una fracción de aminoácidos puede ser invariable (+I) o asignando a cada sitio una probabilidad a pertenecer a una categoría de tasa (+G), indicando que no todas las posiciones tienen la misma tasa de sustitución; adicionalmente, puede ser tomada en cuenta la frecuencia de los aminoácidos observados (+F) (Abascal *et al.*, 2005).

La selección del modelo puede verse como una forma de identificar aquel modelo que, entre un conjunto de candidatos, se acerque más a la realidad (Abascal *et al.*, 2005). Actualmente uno de los mayores retos en la filogenética es la elección exacta del modelo de sustitución de nucleótidos o de aminoácidos entre un conjunto de modelos con el fin de inferir una filogenia exacta. Felsenstein fue el primero en mostrar en simulaciones que los modelos simples que subestiman la sustitución múltiple pueden resultar en inconsistencias durante la estimación de la filogenia, en ciertas situaciones (Keane *et al.*, 2006). Existen diversos programas para la elección del modelo de sustitución para nucleótidos, los cuales se basan en la función de Verosimilitud. En contraste, están

disponibles muy pocos programas para la elección de modelos de sustitución de proteínas actualmente como lo son ModelGenerator y ProtTest. Este último programa es el más popular y su tarea es buscar el mejor modelo de reemplazamiento de aminoácidos para un alineamiento de proteínas dado y que se deriva del programa ModelTest (el cual sirve para elegir el modelo de sustitución para secuencias de nucleótidos). Dado un alineamiento de proteínas y la topología de un árbol el programa calcula la verosimilitud (L) bajo cada uno de los modelos candidatos y estima los parámetros del modelo, la versión 1.2 implementa 64 modelos empíricos: las 8 matrices WAG, mtREV, Dayhoff, JTT, VT, Blosum62, CpREV y RtREV bajo +F, +G, +I y sus combinaciones. Para cada uno de los modelos, la topología del árbol puede fijarse (proporcionada por el usuario o calculada mediante el programa BIONJ, el cual emplea una versión mejorada del algoritmo NJ de Saitou y Nei; sin embargo, BIONJ usa un modelo simple de primer orden de las varianzas y covarianzas de las distancias evolutivas estimadas (Gascuel, 1997)), seleccionando como la estrategia de optimización “*fast*” en la consola de ProtTest, o puede optimizarse (la longitud de ramas y los parámetros del modelo, para cada uno de los modelos) para asegurar que se logró la ML. La optimización completa puede realizarse en ProtTest, cuando se selecciona como estrategia de optimización la opción “*slow*”. Después de esto, el usuario puede elegir un estrategia de selección del modelo como los son el criterio de información de Akaike (AIC), la corrección del criterio de información de Akaike (AICc) o el criterio de información Bayesiano (BIC), y obtener un ordenamiento de los modelos, los parámetros promedio estimados del modelo o mediciones de parámetros importantes (Abascal *et al.*, 2005).

En la elección del modelo es necesario encontrar el balance entre la exactitud y la simplicidad del modelo. Akaike encontró una relación simple entre la verosimilitud (L) y el número de parámetros (K) para estimar la distancia esperada de un modelo dado desde el verdadero:

$$AIC = -2\ln L + 2K \text{ (Abascal } et al., 2005)$$

Cuando el tamaño de la muestra (n) es más pequeño cuando se le compara con el número de parámetros, el AIC podría no ser exacto y por ello es necesario usar la corrección de AIC (AICc):

$$AICc = AIC + (2K(K+1)/n-K-1) \text{ (Abascal } et al., 2005)$$

Una aproximación diferente pero simple para la elección del modelo es el Criterio de Información Bayesiana (BIC), formulado como:

$$\text{BIC} = -2\ln L + K \log n \text{ (Abascal et al., 2005)}$$

Al igual que con otros métodos de reconstrucciones filogenéticas se cuestiona la exactitud de los árboles obtenidos por máxima verosimilitud. Sin embargo, algunos estudios de simulación sugieren que la máxima verosimilitud es típicamente más exacta (tiene mayor probabilidades de predecir el árbol evolutivo real) y más robusta (esto es, menos sensible a los modelos y las suposiciones incorrectas) que otros métodos de inferencia filogenética (Huelsenbeck y Rannala, 1997).

La inferencia bayesiana de árboles filogenéticos es una innovación reciente que simultáneamente estima y proporciona una medición del árbol y de los parámetros en el modelo evolutivo (Keith, 2008). El campo de la estadística bayesiana está estrechamente relacionado con la máxima verosimilitud (Holder y Lewis, 2003); la principal diferencia teórica entre la inferencia bayesiana y las aproximaciones que emplean verosimilitud es que las inferencias bayesianas incluyen un factor que describe las expectativas previas o anteriores acerca de un problema. Estas expectativas previas o anteriores en las inferencias bayesianas, se conoce como la probabilidad anterior, y se refiere a una distribución de probabilidades de los parámetros, incluyendo el árbol y el modelo, y describe que tan frecuentemente uno puede esperar que los valores sean observados antes de evaluar la evidencia desde los datos (Keith, 2008). La hipótesis óptima es la que maximiza la probabilidad posterior. La probabilidad posterior para una hipótesis es proporcional a la verosimilitud multiplicada por la probabilidad anterior de esa hipótesis (Holder y Lewis, 2003). Las inferencias bayesianas examinan la distribución posterior de los parámetros de interés, como las probabilidades relativas de los diferentes árboles. La probabilidad posterior es una distribución de probabilidad formada como una función de la probabilidad anterior y la verosimilitud, la cual representa la información contenida en los datos (Keith, 2008). Las principales ventajas de las inferencias bayesianas es que permite la estimación del árbol y de los parámetros al mismo tiempo que proporciona mediciones de soporte, y permite la implementación de modelos complejos (demasiados

parámetros) de evolución de secuencias, los cuales no pueden emplearse en métodos basados en máxima verosimilitud, debido a que sería muy lento el cálculo de los parámetros (Holder y Lewis, 2003).

Un problema con los métodos basados sobre el principio de optimización es la enorme cantidad de tiempo de cómputo que se requiere cuando el número de secuencias es grande, especialmente con los métodos de máxima verosimilitud. Por esta razón hoy en día se usan algoritmos de búsqueda, que tratan de aproximarse al árbol óptimo en forma heurística (Takahashi y Nei, 2000).

Un árbol filogenético es una hipótesis de las relaciones evolutivas, pero como en el presente no podemos observar directamente el pasado, no se puede probar que tanto el árbol propuesto refleja los verdaderos eventos evolutivos. Sin embargo, hay distintas formas en las cuales la robustez de los datos con los que se construyen los árboles puede probarse. Los cuantificadores que se usan más comúnmente son los valores de soporte de *bootstrap* y el *jack-knife* (Harrison y Langdale, 2006). En el *bootstrap*, la matriz de datos originales se vuelve a muestrear, con reemplazamiento para producir pseudorrelicas de los datos, y se construye un árbol empleando cada uno estos datos. El método de *bootstrap* ofrece una medición de que partes del árbol se soportan débilmente soportadas. Una agrupación que está representada en un bajo porcentaje de las réplicas de *bootstrap* es sensible a la combinación exacta de sitios que se secuenciaron. Esto implica que si otro conjunto de datos se colectara, hay una alta probabilidad de que el grupo pudiera no recobrase. Los valores de *jack-knife* muestran el porcentaje de veces que un clado aparece cuando un porcentaje de los caracteres específicos se remueven aleatoriamente desde los datos y el análisis se realiza nuevamente (Harrison y Langdale, 2006). Algunos trabajos han concluido que la proporción de *bootstrap* es una medida conservadora del soporte, y un valor de 70% indica un fuerte apoyo para el grupo. Es también importante tener en mente que la proporción de *bootstrap* ayuda a predecir si los mismos resultados se obtendrían si se colectan más datos, y no si el resultado es correcto (Holder y Lewis, 2003). Como una medición de soporte estadístico no tradicional encontramos la probabilidad posterior, la cual se emplea específicamente en las inferencias bayesianas. Este método al igual que en el *bootstrap* se genera una muestra

de árboles, mediante la metodología *Markov chain Monte Carlo* (mcmc). A partir de un árbol inicial se propone un nuevo árbol, la aceptación o rechazo de este nuevo árbol involucra una elección que satisfaga distintas condiciones, en el caso de las inferencias bayesianas, esta condición es la probabilidad posterior (Holder y Lewis, 2003). Dentro de la muestra de árboles, aquellos grupos que aparecen como monofileticos y que estén altamente representados tienen una alta probabilidad de que esta suposición sea correcta. Por ejemplo, si de una muestra de 10^5 árboles un clado se recobra en 975676 de ellos, dicho clado tiene una probabilidad posterior de ≈ 0.98 .

El enraizamiento de un árbol filogenético nos permite darle una polaridad o dirección e inferir el orden de los eventos de cambio. Los métodos filogenéticos dan como resultado solamente diagramas de ramificación, a partir de los cuales, en algunas ocasiones, es difícil establecer el orden de los cambios, lo cual es fundamental cuando se investiga los procesos evolutivos; sin embargo, no siempre es necesario enraizar los árboles (Harrison y Langdale, 2006).

El método más empleado para enraizar los árboles filogenéticos de datos moleculares es el enraizamiento mediante un grupo externo, que es aquel que está más cercanamente relacionado con el grupo de interés. Las diferencias entre ambos grupos se emplean para inferir la dirección de los cambios. El grupo externo puede seleccionarse basándose en el conocimiento previo que se tenga del grupo de interés, o podría aparecer durante el alineamiento. En las familias de genes, un grupo externo adecuado debe compartir al menos un dominio con el grupo de interés; sin embargo, no es trivial encontrar un grupo externo apropiado, porque dentro de las secuencias candidatas, los dominios que pueden ser adecuadamente alineados para el análisis filogenético, pueden no proporcionar suficiente variación y otros dominios pueden variar demasiado. Por otra parte, también se puede emplear la duplicación de genes para llevar a cabo el enraizamiento, donde las secuencias de un clado de genes se usan como la raíz del otro. El enraizamiento empleando genes duplicados tiene la ventaja de que pueden revelar relaciones inesperadas entre especies o genes en los clados principales, donde previamente se ha enraizado con ambigüedad (Harrison y Langdale, 2006).

Como una alternativa a los métodos anteriores, se desarrolló el método de enraizamiento de árboles reconciliados. La reconciliación es el proceso de construcción de un mapa entre un árbol de una familia de genes y un árbol de las especies. Bajo un modelo de parsimonia duplicación-pérdida, este mapa puede usarse para inferir las duplicaciones y las pérdidas en la historia de la familia de genes, además de los linajes de especies en los cuales estos eventos ocurrieron. La reconciliación es un método esencial en filogenética molecular que se utiliza ampliamente en aplicaciones evolutivas, en la medicina, en la biología del desarrollo y las ciencias de las plantas. La reconciliación, después de la reconstrucción filogenética, es la aproximación más segura para identificar ortólogos para la predicción de función, la anotación de genes y la identificación de blancos de medicamentos. La reconciliación se usa para correlacionar duplicaciones específicas con el surgimiento de nuevas funciones celulares o características morfológicas, proporcionando pistas sobre la función de nuevos genes descubiertos. La reconciliación es también el núcleo de un problema relacionado, pero más complejo: la inferencia de un árbol de especies desde algunos árboles de genes (Vernot et al., 2007).

El método de árboles reconciliados, en una explicación sencilla, consiste en hacer coincidir el árbol de genes con el árbol de especies, el cual se cree que refleja correctamente las relaciones evolutivas, cuando el árbol de genes presenta discordancias con este último. Este método parte del supuesto que ambos árboles, el de genes y el de las especies, son correctos, sin embargo, la presencia de discordancias entre ellos se podría explicar mediante la postulación de duplicación de genes y de la pérdida de estos (Page, 2000).

Supongamos que tenemos un árbol con 4 especies que representa las relaciones de ancestría-descendencia y por otra parte tenemos un árbol con 4 genes de las 4 especies como se muestra en la figura 3a); sin embargo, notamos que hay discordancias cuando los comparamos. Para poder resolver estas discordancias tenemos que postular eventos de duplicación en el ancestro común de las especies 2, 3 y 4, y de pérdida, por un evento de muestreo ostensible, figura 3b), el cual es la aparente ausencia de un gen en los descendientes de un ancestro que presentaba el gen, estos eventos de muestreo ostensible implican la pérdida del gen, pero también pueden reflejar fallas técnicas en la detección de la copia del gen en la especie (Page, 2000).

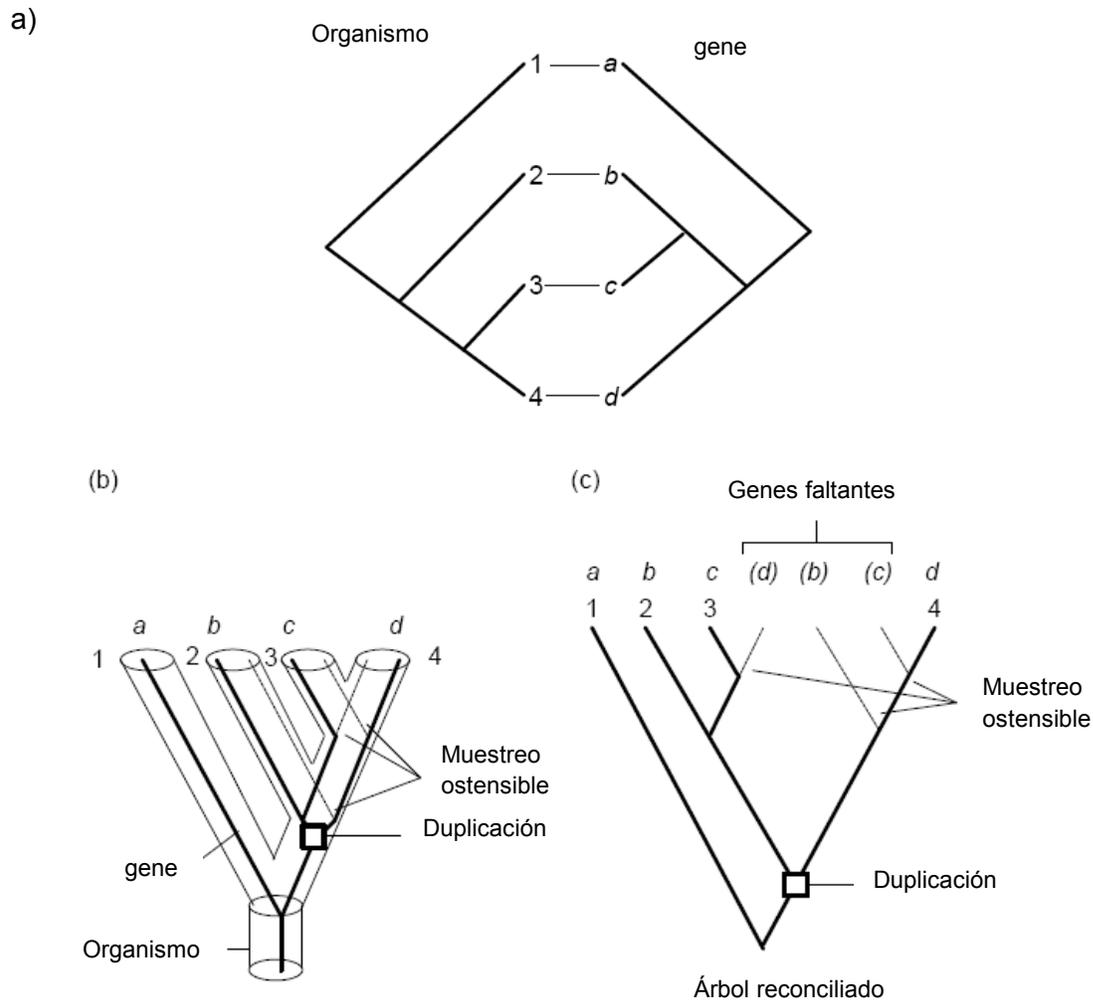


Figura 3. a) Discordancias entre el árbol de las especies y el árbol de genes; b) postulación de eventos de duplicación y de muestreo ostensible para explicar las discordancias entre los árboles mostrados en la a); c) árbol reconciliado, que explica la historia evolutiva del gen (Page, 2000).

2.3 Detección de Selección Natural a Nivel Molecular

La selección natural es una de las principales fuerzas evolutivas que moldea la variación genética (cambios en el genoma) mediante la eliminación o el mantenimiento de ésta. Se consideran dos tipos de selección natural más importantes a nivel molecular, la selección positiva que es el mecanismo mediante el cual las nuevas mutaciones se mantienen en las poblaciones debido a que aumentan la adecuación, y la selección negativa en la que las nuevas mutaciones que poseen una adecuación menor se eliminan. La selección positiva promueve la diversidad genética, mientras que la selección negativa disminuye la diversidad mediante la eliminación de las variaciones. A nivel de secuencias codificantes

de proteínas, las mutaciones que aumentan la adecuación son aquellas que mejoran la función u originan una nueva función, y las que disminuyen la adecuación son las que afectan la función o la estructura.

La inferencia de la presión de selección sobre los sitios de aminoácidos individuales proporciona una importante aproximación para estudiar los mecanismos de la evolución y de la función de proteínas. Se han desarrollado distintos métodos, que pueden detectar sitios funcionalmente importantes basándose en la conservación evolutiva. Sin embargo, altos niveles de variabilidad también tienen importante significado funcional. La evidencia de que cada uno de los residuos experimente selección positiva puede inferirse mediante la observación de que las tasas de sustitución no-sinónimas (dN) es más alta que la tasas de sustitución sinónimas (dS) de nucleótidos en un gene que codifica para una proteína (Ouyang y Liang, 2007). Tomando en cuenta que las mutaciones sinónimas (silenciosas) son en gran parte invisibles para la selección natural, y que las mutaciones no-sinónimas (reemplazamiento de aminoácidos) pueden estar bajo fuerte presión selectiva, la comparación de las tasas de fijación de aquellos dos tipos de mutaciones proporcionan una poderosa herramienta para entender el mecanismo de evolución (Yang y Nielsen, 2000). La proporción entre las tasas no sinónimas y sinónimas ($\omega = dN/dS$) se ha tomado como una medición estándar de presión de selección (Yang y Nielsen, 2002) de la dirección y de la magnitud de la selección sobre el cambio del aminoácidos (Yang, 2007). Se espera que $\omega = dN/dS = 1$ si no hay selección tipo Darwiniana que esté actuando sobre las secuencias de DNA; $\omega < 1$ (selección en contra de nuevas mutaciones) si hay selección negativa; y $\omega > 1$ (selección para nuevas mutaciones) si hay selección positiva (Ouyang y Liang, 2007); una proporción de ω mayor que uno significa que las mutaciones no-sinónimas ofrecen ventajas de adecuación y se fijan en la población a una tasa más alta que las mutaciones sinónimas (Yang y Nielsen, 2002). El problema es la estimación de las sustituciones (dS) sinónimas y (dN) no-sinónimas por sitio entre dos secuencias (Yang y Nielsen, 2000). Se han propuestos distintos métodos para detectar la presencia de la selección positiva en un alineamiento de secuencias homólogas de genes codificantes de proteínas, desde el conteo de diferencias sinónimas y no-sinónimas observadas entre pares de secuencias, hasta las técnicas más sofisticadas que toman en cuenta las relaciones filogenéticas entre algunas secuencias y permiten una prueba estadística de los resultados. Estos últimos métodos extraen mas información mediante la

consideración de las relaciones evolutivas entre las secuencias analizadas pero difieren marcadamente en sus aproximaciones (Massingham y Goldman, 2004).

En las décadas pasadas, se ha sugerido un número de métodos intuitivos para esta estimación (Yang y Nielsen, 2000). dN y dS pueden estimarse basándose sobre los métodos de aproximación, los cuales típicamente involucran tres pasos esenciales: 1) conteo del número de de sitios sinónimos(s) y no-sinónimos (N) entre las secuencias blanco, 2) conteo del número de sustituciones sinónimas (S) y no-sinónimas (N) entre dos secuencias ortólogas, y 3) calculando el número de sustituciones sinónimas (dS) y no-sinónimas (dN) por sitio después de la corrección para sustituciones múltiples. La mayoría de los métodos asume simplificaciones de rutas de sustitución de nucleótidos e involucran un tratamiento de datos *ad hoc* que no están bien justificados (Wang *et al.*, 2009). Dentro de los métodos más recientes se encuentran los de máxima verosimilitud (ML) que emplean los modelos de sustitución de codones desarrollado por Goldman y Yang (1994) o por Muse y Gaut (1994). Estos modelos de codones son intuitivamente atractivos en su uso de la relación de tasas de sustitución no-sinónimas a sinónimas ($\omega = dN/dS$) para definir el tipo y la fuerza de la selección. Los métodos ML son extremadamente flexibles, ya que permiten que múltiples modelos de evolución de secuencia puedan ajustarse a los datos mientras que incorporan una variedad de parámetros de sustitución, los cuales pueden variar dentro de una filogenia o a lo largo de una secuencia. Los modelos de codón que permiten presiones de selección heterogéneas entre sitios han llamado la atención por su poder para detectar selección y su potencial utilidad en la predicción de qué sitios están bajo selección (Aagaard y Phillips, 2005). Una importante mejora en los métodos de aproximación, dentro de un marco de máxima verosimilitud, es el propuesto por Yang y Nielsen (2000), los cuales consideran 2 características importantes sobre la evolución del DNA, para el cálculo de las tasas de sustitución no-sinónimas y sinónimas, que son los sesgos en las tasas de transición/transversión y en la frecuencias de los nucleótidos.

Por otra parte se han desarrollado modelos alternos que permiten que, durante el proceso de evolución, varíe la relación de las tasas de sustitución no-sinónima/sinónimas (ω), no sólo entre sitios, sino también entre las ramas o en ambos. En los modelos sitio-

específico, se trata la relación ω para algún sitio (codón) en el gen como una variable aleatoria obtenida de una distribución estadística, y permite que ω varíe entre codones (Yang, 2007). Estos modelos permiten variar a la proporción ω entre sitios pero no entre linajes; y la selección positiva se detecta en sitios individuales solo si el promedio de dN es mayor en todos los linajes que el promedio de dS (Yang y Nielsen, 2002); a diferencia con el modelo anterior, los modelos de rama o linaje-específico pueden usarse para detectar eventos de selección positiva cuando está actuando sobre linajes particulares (Yang, 2007). En estos modelos se permite variar a ω entre los linajes; se asume que no hay variación de ω entre sitios y como resultado se puede detectar selección positiva para linajes sólo si el promedio de dN es mayor que el promedio de dS de una rama específica (Yang y Nielsen, 2002). Los modelos de ramas son útiles para detectar selección positiva después de la duplicación de genes, si una copia de las duplicadas pudo adquirir una nueva función y pudo haber evolucionado con tasas aceleradas (Yang, 2007).

Como una alternativa a los métodos anteriores, los cuales parecen ser muy estrictos al permitir variar ω sólo a través de sitios o de ramas, pero no en ambos, lo cual les puede restar poder para detectar sitios bajo selección positiva, bajo ciertas circunstancias. Existen los modelos sitio-rama para detectar selección positiva que solo afecte unos pocos sitios sobre linajes pre-especificados (Yang, 2007). En este método se asume que la filogenia se conoce o se estima independientemente, así como se conocen las ramas que pueden estar bajo selección (Yang y Nielsen, 2002). Las ramas bajo prueba para selección positiva se llaman ramas *foreground*, mientras las demás ramas se les denominan como ramas *background* (Zhang *et al.*, 2005; Yang, 2007; Yang y Nielsen, 2002). La prueba sitio-rama requiere de una especificación *a priori* de las ramas *foreground*. Cuando múltiples ramas sobre el árbol se prueban para selección positiva con el mismo conjunto de datos, se requiere una corrección para pruebas múltiples. Un procedimiento simple pero conservador es la corrección de Bonferroni, lo que significa que la prueba individual de cualquiera de las ramas se considera significativa en el nivel de α solo si el valor p es $< \alpha/m$, donde m es el número de ramas que se prueban con los mismos datos (Yang, 2007).

Sin embargo, una desventaja de los métodos de sitio-rama es que se tiene que especificar *a priori* los linajes que se consideran que están bajo selección positiva (ramas *foreground*), lo cual implica, a su vez, la asignación de un modelo de evolución, del cual

no se tiene la certeza de que sea el mejor modelo. Por otra parte, en la ausencia de una hipótesis biológica, la selección de las ramas a ser probadas para la presencia de selección positiva, ramas *foreground*, podría llegar a ser de una forma arbitraria, lo cual conlleva el riesgo de hacer suposiciones erróneas de presencia de selección en estas ramas. Con respecto a que la relación de tasas de sustitución no-sinónimas/sinónimas (ω) es la misma para todas las ramas *background* en la filogenia, suena poco realista.

Como una alternativa a los modelos anteriores de detección de presión de selección, los cuales presentan ciertas restricciones y limitaciones sobre todo en la elección del modelo de evolución, Kosakovsky-Pond y Frost (2005a) desarrollaron un nuevo método llamado GA-branch, el cual realiza la detección de presión de selección a nivel de ramas mediante la asignación de un valor de dN/dS a cada una de las ramas en el árbol filogenético y tiene como principal ventaja que el modelo que se selecciona es aquel que mejor se ajusta estadísticamente a los datos. Como principal desventaja del método GA-branch, es que actualmente solo puede detectar presión de selección a nivel de ramas.

2.4 Antecedentes

Los proyectos de secuenciación de genomas han revelado la existencia de familias de genes para cada uno de las subunidades del factor eIF4F, en numerosas especies eucarióticas (Hernández y Vazquez-Pianzola, 2005); eIF4E y todas las proteínas que presentan cierta similitud de secuencia con esta proteína, forman la familia eIF4E. Dados algunos roles críticos que juega eIF4E en la célula, no es sorprendente que la secuencia primaria de aminoácidos esté altamente conservada (Goodfellow y Roberts, 2008); aunque las regiones amino-terminal y carboxilo-terminal son considerablemente variables tanto en longitud como en secuencia (Joshi *et al.*, 2004). Los alineamientos de las secuencias completas de los aminoácidos de los factores de iniciación de la traducción de *Mus musculus* (eIF4E-1), *Danio rerio* (eIF4E-1^a), *Drosophila melanogaster* (eIF4E-1), *Triticum aestivum* (eIF4E y eIF(iso)4E), *Schizosaccharomyces pombe* (eIF4E1) y *Saccharomyces cerevisiae* (eIF4E), sugieren la presencia de una región *core* conservada. Esta región se extiende aproximadamente por 160-170 residuos desde la His-37 a la His-200 de eIF4E *Homo sapiens* y *Mus musculus* (figura 4). La secuencia consenso de la

región *core* conservada porta las suficientes características para definir estas proteínas como miembros de la familia eIF4E; los residuos aromáticos de triptófano, fenilalanina e histidina muestran un patrón distintivo, resumido por:

His(X5)Trp(X2)Trp(X8-12)Trp(X9)Phe(X5)PheTrp(X20)Phe(X7)Trp(X10)Trp(X9-12)Trp(X34-35)Trp(X32-34)His (Joshi *et al.*, 2005).

Esta secuencia consenso indica que la región *core* de eIF4E, inicia y termina con un residuo de histidina; los 8 residuos de triptófanos y los 2 residuos de fenilalanina, característicos de estas proteínas, cada uno está separado por un número variable de residuos de aminoácidos que pueden ser cualquiera de los 20 residuos (en la secuencia consenso anterior esto es representando mediante la X seguido del numero de residuos).

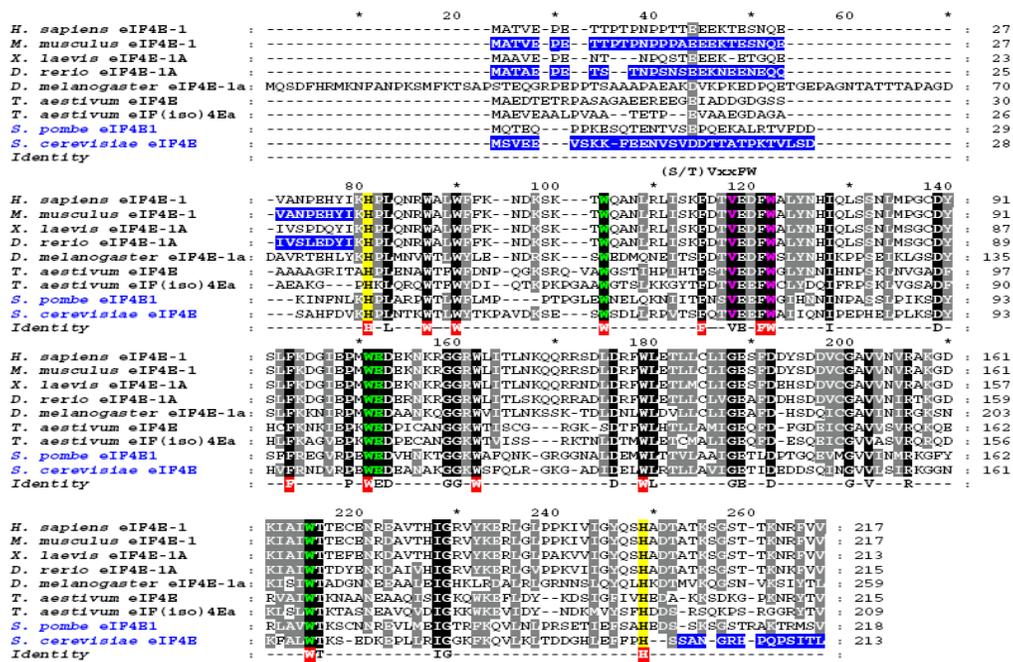


Figura 4. Alineamiento de secuencias de aminoácidos de miembros de la familia eIF4E, en las cajas rojas se muestran residuos aromáticos idénticos en todas las secuencias, mientras que las cajas amarillas son los residuos de histidina que delimitan el *core*; los residuos en verde indican los triptófanos 56, 102 y 166 que son residuos que interactúan con el *cap*. Los residuos en purpura son aquellos que interactúan con eIF4G y 4EBP's (Joshi *et al.*, 2005).

La clase estructural I incluye a los miembros de los subgrupos 1, 2 y 3; la comparación de secuencias de residuos de aminoácidos que representan la región *core* de miembros de los subgrupos 1, 2 y 3 revela que comparten aproximadamente 35 a 40% de identidad y de 60 a 65% de similitud; los miembros de esta clase poseen residuos de Trp equivalentes a Trp-43 y Trp-56 de eIF4E-1 de *Homo sapiens* e incluyen eIF4E-1 de Metazoos, eIF4E(p26) y eIF(iso)4E(p28) de plantas, y eIF4E de Hongos (Joshi *et al.*, 2005). Los miembros pertenecientes a esta clase se expresan ubicuamente y son responsables de llevar a cabo la traducción general dependiente de *cap* (Joshi *et al.*, 2005).

Los miembros de la clase estructural II incluyen a eIF4E-2 de Metazoa (subgrupo 4); nCBP de Viridiplantae (subgrupo 5) y eIF4E-2 de hongos patógenos del sub-phylum Pezizomycotina, incluyendo a *Coccidioides posadasii* y *Sclerotinia scleroiorum* (subgrupo 6). La comparación de secuencias de las regiones *core* de los miembros de esta clase muestra una identidad de aproximadamente 50% y una similitud de 70 a 80%; así como también presentan una identidad de 30 a 35% y una similitud de 60 a 65% con los miembros de la clase I. Los miembros de la clase II difieren de la clase I en que poseen un residuo hidrofóbico como Tyr, Phe o Leu pero no Trp, en la posición equivalente a Trp-43 de eIF4E de *H. sapiens*; así como también contienen una sustitución en la posición equivalente a Trp-56 por Tyr o Phe. Algunos estudios han demostrado que proteínas pertenecientes a la clase II, eIF4E-2A (4EHP) de *H. sapiens*, eIF4E-2 de *M. musculus*, IFE-4 de *C. elegans*, eIF4E-2 en *D. melanogaster* y nCBP de *Arabidopsis thaliana* todas se unen a la estructura *cap*. Sin embargo, nCBP de *A. thaliana* difiere de eIF4E-2A de *H. sapiens*, *M. musculus* y eIF4E-2 de *D. melanogaster*, ya que nCBP puede interactuar con eIF4G y participar en la traducción (Joshi *et al.*, 2005). El rol fisiológico de eIF4E-2A o 4EHP en *H. sapiens* aun no está claro. Sin embargo, se ha reportado que no se une a eIF4G y consecuentemente no puede llevar a cabo las mismas funciones que eIF4E. Recientemente se ha descrito una función para 4EHP para el homólogo en *Drosophila* d4EHP, donde el desarrollo correcto del embrión requiere de una distribución asimétrica de la proteína *caudal* que se logra mediante la regulación a nivel traduccional e involucra a d4EHP en un complejo con la proteína *bicoid*. La proteína d4EHP se une tanto a la

estructura *cap* 5' del RNAm de *caudal*, como a *bicoid* que, a su vez, interactúa con el extremo 3' del RNAm, previniendo que eIF4F se una y la traducción pueda llevarse a cabo. La expresión de *caudal*, por ende, depende de la concentración local del complejo d4EHP-Bicoid (Rosettani *et al.*, 2007). Debido a que eIF4E-2A de los Metazoos no puede participar en la síntesis de proteínas por su inhabilidad para interactuar con eIF4G, se sugiere que tiene un rol regulatorio (Joshi *et al.*, 2005).

En la clase estructural III se encuentran los miembros de la familia de eIF4E, pertenecientes al subgrupo 7, y comparten un 25 a 30% de identidad y 45 a 55% de similitud con los miembros de la clase I y clase II. Esta clase presenta un residuo de Cys o Tyr en la posición equivalente al Trp-56 de eIF4E de *H. sapiens*. Debido a que no presentan el residuo de Trp 56, el cual participa directamente en la unión a la estructura *cap*, la habilidad para interactuar con ésta puede no presentarse. Los estudios con eIF4E-3 de *M. musculus* han mostrado que la proteína interactúa débilmente con la estructura *cap in vitro*. Sin embargo, eIF4E-3 de *M. musculus* es incapaz de rescatar el crecimiento de *S. cerevisiae* carente de un gen funcional de eIF4E. La interacción débil de eIF4E-3 de *M. musculus* sugiere que esta proteína puede estar involucrada en el secuestro de eIF4G que resulta en la inhibición de la traducción dependiente de *cap* (Joshi *et al.*, 2005).

Los miembros del subgrupo 8 poseen aproximadamente un 20% de identidad y un 40% de similitud con respecto a la región *core* de los miembros de la clase I, II y III; presentan un residuo de aminoácido de Trp o Tyr equivalente al Trp 56 de eIF4E de *H. sapiens* y un Trp equivalente al residuo de Trp 43 de eIF4E de *H. sapiens*. Sin embargo, los miembros de este subgrupo también presentan extensiones de 12 a 15 residuos de aminoácidos entre los residuos equivalentes a los Trp 73 y Trp 102 de eIF4E-1 de *H. sapiens*, y de 4 a 9 entre los residuos equivalentes a los Trp 102 y Trp 166 (Joshi *et al.*, 2005).

Aun no está claro a qué se deba la presencia ni cuál es la función de las múltiples isoformas de eIF4E. Esta multiplicidad podría reflejar una simple redundancia funcional, pero también sugiere un rol más complejo para las isoformas de eIF4E. Esta posibilidad

se ha explorado en *C. elegans*, en el cual la existencia de distintas proteínas parecidas a eIF4E se atribuyó a la necesidad de reclutar RNAm con diferentes *cap* al complejo de iniciación (Li y Wang, 2005). *C. elegans* presenta 5 isoformas de eIF4E, tres de ellas (IFE-1, IFE-2 y EIF-5) se unen tanto a *cap* 7-metilguanosina como a 2,2,7-trimetilguanosina en el extremo 5' de los RNAm de *C. elegans* (Ptushkina *et al.*, 2001). Las isoformas de *C. elegans* tienen diferentes afinidades por las estructuras *cap* trimetiladas (Jankowska-Anyszka *et al.*, 1998). La presencia de eIF4E específicos que se unan a *caps* trimetilados es esencial en un organismo en el cual aproximadamente el 70% de sus RNAm tienen esta estructura. Una cuarta isoforma (IFE-3) es específica para estructuras monometiladas y es esencial, mientras que una quinta isoforma (IFE-4), que también es específica de 7-metilguanosina, no es esencial. IFE-4 está más cercanamente relacionada (más que las otras isoformas de *C. elegans*) a las proteínas de la clase II encontradas en *A. tahaliana* (nCBP) y vertebrados (4EHP) (Ptushkina *et al.*, 2001). En otro caso, una de las dos isoformas de eIF4E en el pez cebra, eIF4E-1B, no puede unirse a la estructura *cap* 7-metilguanosina y aparentemente no participa en la iniciación de la traducción (Li y Wang, 2005). A pesar de ello comparte un 66% de identidad con el prototipo eIF4E y posee todos los residuos que se requieren para unirse a la estructura *cap*, eIF4G y 4E-BPs. Además, datos de los análisis de expresión en tejidos del pez cebra, de cultivos celulares y en diferentes estados de desarrollo sugieren que la presencia de eIF4E-1B no es esencial para la sobrevivencia (Robalino *et al.*, 2004). Sin embargo, en la mayoría de los sistemas, el significado biológico de las múltiples isoformas de eIF4E permanece incierto.

3. Hipótesis y Objetivos

Las mejoras metodológicas aunadas a los avances tecnológicos, han permitido un crecimiento exponencial en la cantidad de datos de secuencias moleculares. Sin embargo, nuestro conocimiento funcional y estructural de las proteínas que se deducen de esas secuencias no ha crecido de la misma manera. Un intento para resolver este problema es mediante los algoritmos de predicción y aproximaciones de comparación, que también permiten agrupar las nuevas secuencias dentro de familias de genes y proteínas. La familia del factor de inicio de la traducción eIF4E, agrupa a las secuencias

moleculares que presentan cierta similitud de secuencia, que a su vez se han clasificado dentro de tres clases estructurales, y posiblemente no compartan la misma función, de acuerdo a Joshi *et al.* (2005). Con base a esto la hipótesis de trabajo que se plantea es la siguiente:

No conocemos del todo las funciones de cada clase de eIF4E, pero si encontramos evidencia de evolución adaptativa en el origen de cada clase, podremos concluir que estos eventos están asociados a la adquisición de funciones diferentes.

Por otra parte el objetivo general de este trabajo es:

La detección de sitios bajo selección natural que puedan aportar información sobre las posibles funciones de las múltiples isoformas de eIF4E.

En tanto que los objetivos particulares derivados de lo anterior son los siguientes:

Revisión y refinamiento de la filogenia de la familia de proteínas eIF4E, empleando métodos robustos de reconstrucción de filogenias (Máxima Verosimilitud).

Detección de sitios, en las secuencias de las proteínas de los miembros de las 3 clases estructurales, bajo selección natural.

Detección de sitios bajo evolución adaptativa en las diferentes isoformas de plantas, eIF(iso)4E y nCBP.

4. Material y Métodos

Los métodos de reconstrucción de las filogenias moleculares y los análisis de detección de presencia de selección natural requieren datos de secuencias moleculares; para este trabajo, se tomó el alineamiento de las proteínas de la familia del factor de inicio de la traducción eIF4E, en formato fasta de la base de datos de Pfam.

<http://pfam.sanger.ac.uk/family/PF01652>.

Para el análisis de detección de presencia de selección se empleó el alineamiento de las secuencias en nucleótidos que correspondían a los CDS (coding sequences) de las secuencias de aminoácidos. Para la obtención de los CDS se realizó una búsqueda de las secuencias del alineamiento en aminoácidos, empleando el nombre de entrada (Entry name) de cada secuencia, en la base de datos de UNIPROT (<http://www.uniprot.org>), posteriormente se siguió la liga para llegar a la secuencia del RNAm en la base de datos del GenBank y de ahí descargar el CDS en formato fasta de cada secuencia.

4.1.- Reconstrucción Filogenética de la Familia de Factores de Inicio de la Traducción eIF4E.

La reconstrucción filogenética de la familia de factores de inicio de la traducción eIF4E se llevó a cabo mediante el empleo del método de máxima verosimilitud, debido a que este método es menos sensible a los modelos y a las suposiciones incorrectas (robustez) y tiene mayor probabilidades de predecir el árbol evolutivo real (exactitud) (Huelsenbeck y Rannala, 1997), en comparación con los métodos de distancia (Neighbour-Joining). Las secuencias alineadas (en residuos de aminoácidos) empleadas para la elaboración de la filogenia se tomaron de la base de datos de Pfam.

Posteriormente, las secuencias se clasificaron en las 3 clases estructurales que Joshi planteó en su trabajo del 2005. El criterio que él tomó para hacer su clasificación fue en el tipo de residuos presentes en la posición 43 y 56 (de acuerdo a la numeración de eIF4E1 de *Homo sapiens*); las secuencias pertenecientes a la clase I presentan residuos de triptófano en las posiciones 43 y 56, la clase II presenta tirosina, fenilalanina o leucina en la posición 43 y tirosina o fenilalanina en la posición 56, mientras que la clase III presenta solo sustituciones en la posición 56 por cisteína o tirosina. Seguimos esta clasificación debido a que la nomenclatura de las secuencias en la base de datos no indica si pertenecen a alguna de las clases antes descritas. Por otra parte mediante el empleo del programa de Jalview (Waterhouse, *et al.*, 2009) se eliminaron del alineamiento las secuencias que carecían de la región conservada denominada “core” o estuviera incompleta, así como los extremos aminoterminal y carboxiloterminal, con el objetivo de

trabajar solo con la región “core”, ya que los extremos al presentar una alta variabilidad tanto de secuencia como de longitud en los extremos aminoterminal y carboxiloterminal podrían ocasionar una reconstrucción filogenética errónea al depender de hipótesis implícitas de homología probablemente espurias. El alineamiento resultante de la clasificación y la edición anterior se utilizó para realizar un alineamiento de perfiles empleando el programa ClustalW (Larkin, *et al.*, 2007) y como el perfil al cual se alinearía, el alineamiento referente a la región “core”, tomado de la base de datos de la familia 4E (<http://umbicc3-215.umbi.umd.edu>) (Joshi, *et al.*, 2005).

El alineamiento por perfiles en el cual el perfil 2 (alineamiento de la región core de las secuencias editadas y clasificadas tomadas de Pfam) se alineó al perfil 1 (alineamiento de la región core tomado de la base de datos de la familia 4E) se realizó para corregir errores, si es que los hubiera, en el alineamiento tomado de Pfam.

La elección del modelo de sustitución se realizó con el programa ProtTest (este programa calcula la verosimilitud (L) bajo cada uno de los métodos candidatos y estima los parámetros del moldeo, a partir de un alineamiento de proteínas y de la topología de un árbol (Abascal *et al.*, 2005)). Para la reconstrucción filogenética se empleó el programa PhyML (Guindon y Gascuel, 2003), el cual usa un algoritmo del tipo *hill-climbing* (realiza varias iteraciones hasta alcanzar un óptimo máximo) que ajusta simultáneamente la topología del árbol y la longitud de rama. Este algoritmo inicia con un árbol construido mediante un método de distancia y modifica este árbol para mejorar su verosimilitud en cada una de las iteraciones. Debido a que se ajustan simultáneamente tanto la topología del árbol y la longitud de rama, se requieren, relativamente pocas iteraciones para alcanzar un óptimo.

En la figura 6 se muestra un diagrama de flujo ilustrando la metodología empleada para la reconstrucción filogenética a partir del alineamiento de perfiles.

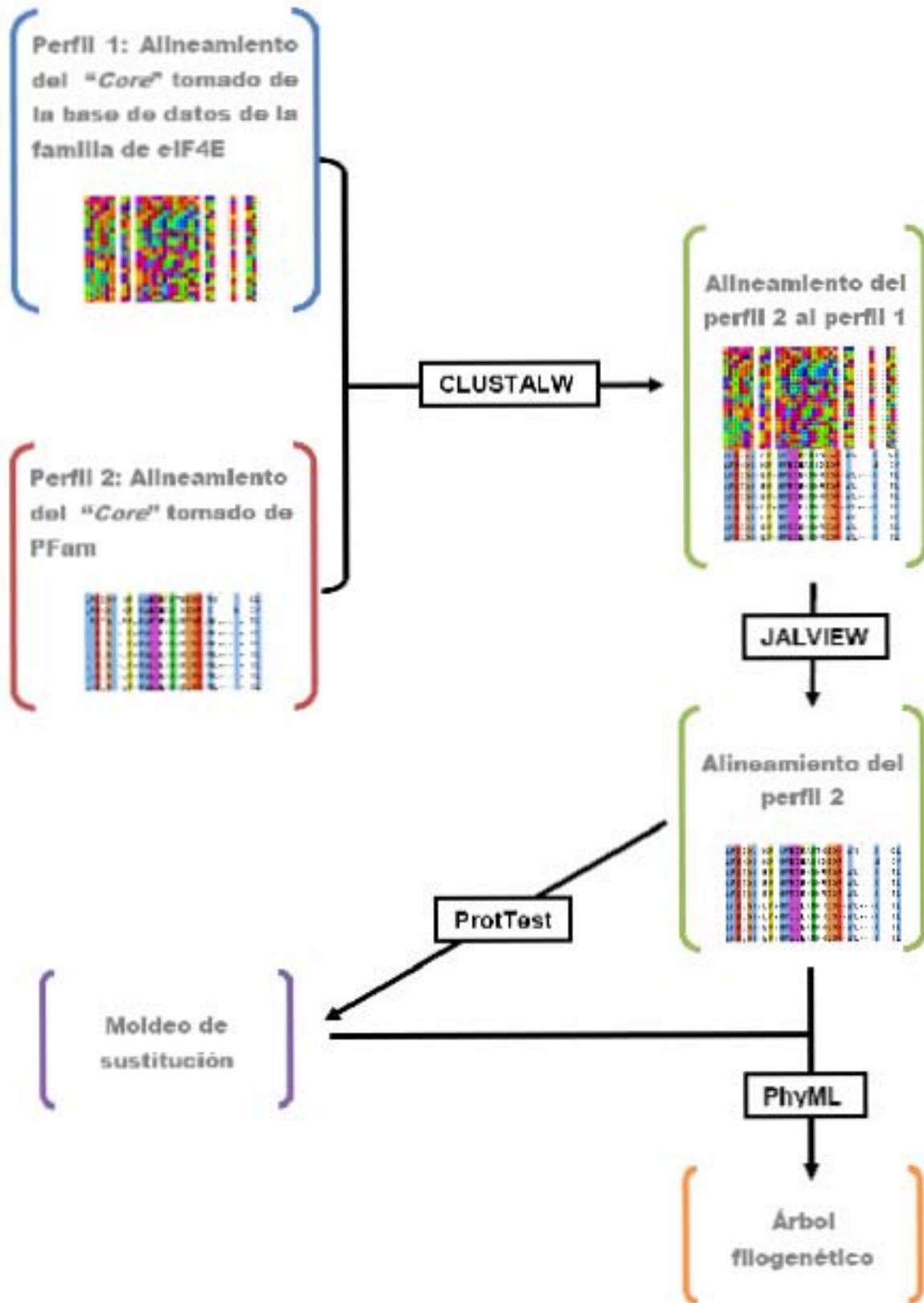


Figura 6. Diagrama de la metodología empleada para la reconstrucción filogenética a partir del alineamiento de perfiles.

4.2 Enraizamiento del árbol filogenético

El enraizamiento de los árboles filogenéticos se llevó a cabo mediante el método de árboles reconciliados, con el programa Notung (Vernot, *et al.*, 2007), el árbol de especies (figura 8) requerido por Notung necesario para llevar a cabo la reconciliación, fue construido con base en los arboles filogenéticos de TREEBASE y de la página web de Tree of life Web Project, así como en la clasificación taxonómica presentada en la página web del NCBI.

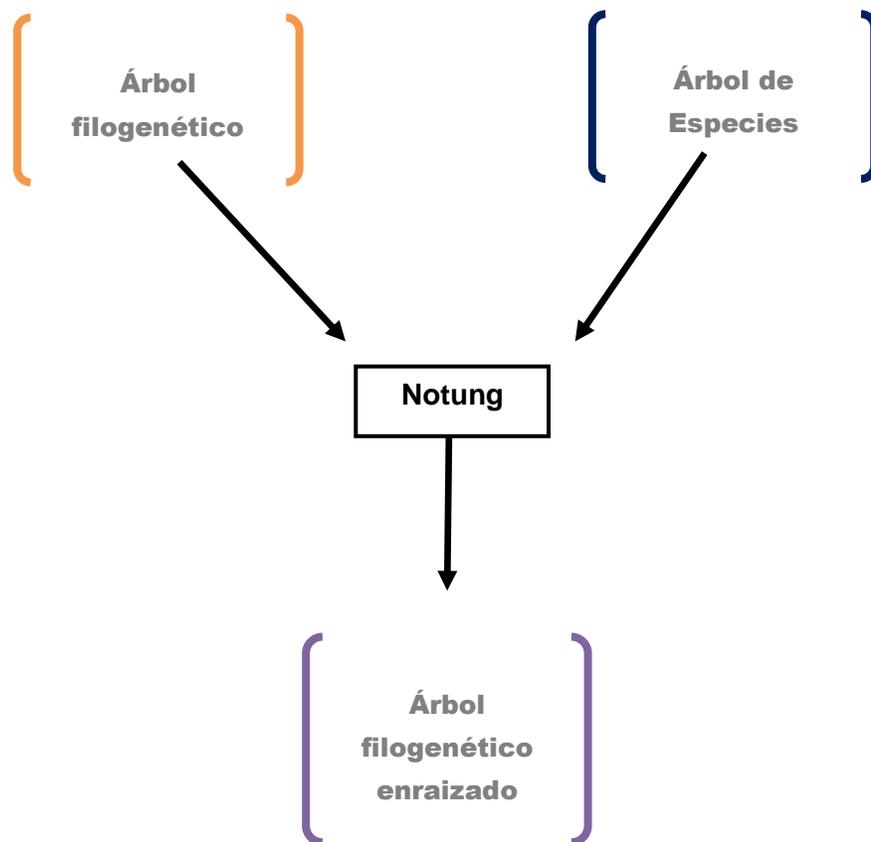


Figura 7. Diagrama de la metodología empleada para enraizar el árbol filogenético, empleando el método de árboles reconciliados

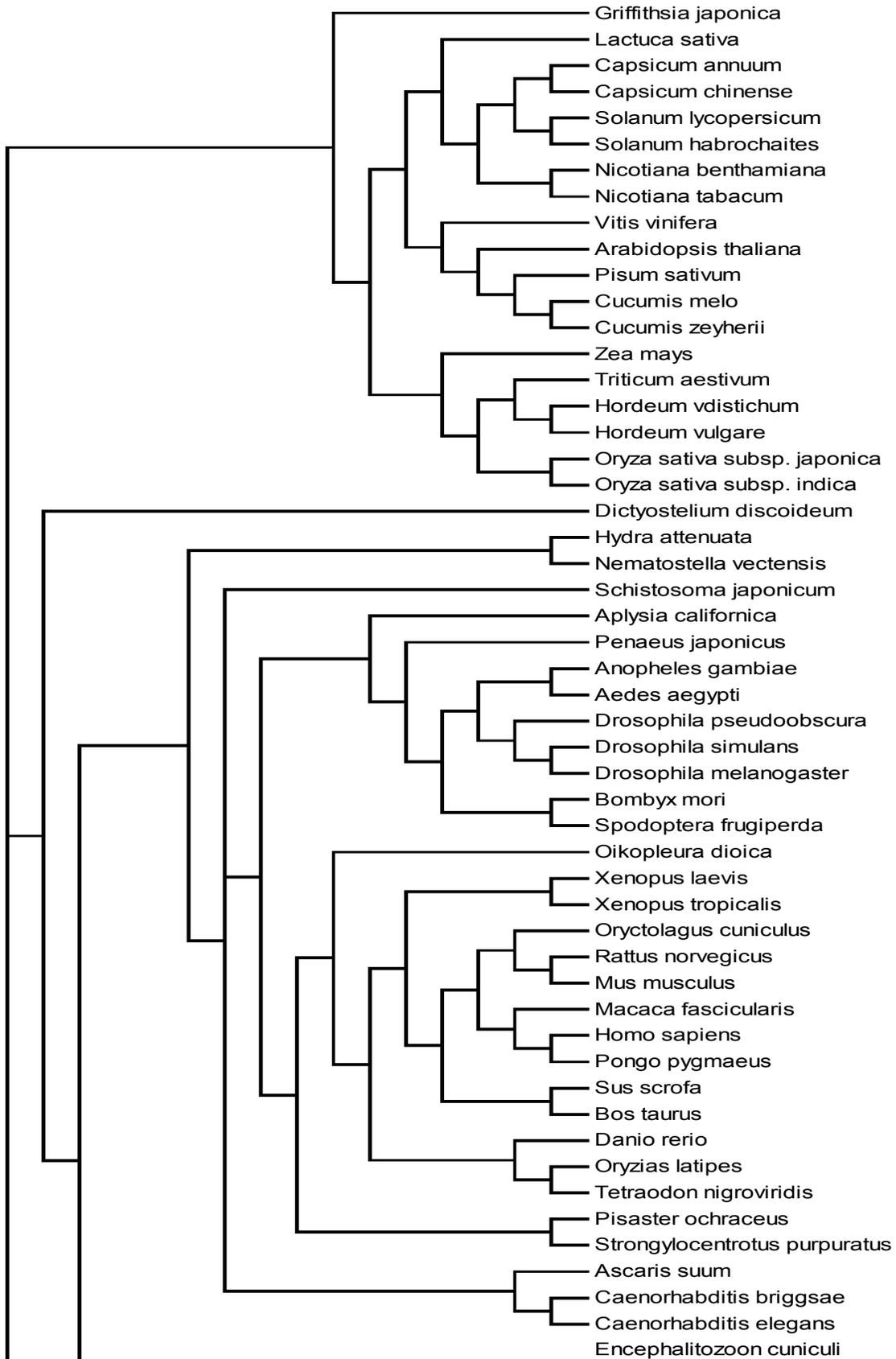




Figura 8. Árbol de especies empleado para la reconciliación.

4.3 Detección de Sitios Bajo Presión de Selección Natural

La detección de sitios bajo presión de Selección Natural se llevó a cabo mediante el uso del alineamiento de secuencias de nucleótidos, que corresponden a los CDS de las secuencias de aminoácidos, presentes en el alineamiento obtenidos de la base de datos de Pfam. Las secuencias de nucleótidos se tomaron del GenBank; se descargaron en formato fasta, y se alinearon mediante el empleo del programa *Tranaling*, el cual alinea las secuencias de nucleótidos dado un alineamiento de aminoácidos. Para ello el alineamiento de aminoácidos fue el mismo utilizado en la reconstrucción filogenética. Posteriormente se construyó un archivo que incluía el alineamiento de nucleótidos, obtenido mediante *Tranaling*, y el árbol filogenético enraizado; este archivo se utilizó para la detección de sitios bajo Selección Natural con el programa HYPHY (Kosakovsky-Pond, *et al.*, 2005) y el método GA-branch (Kosakovsky-Pon y Frost, 2005a), que detecta selección mediante pruebas de ramas mediante la asignación de un valor de dN/dS a cada una de las ramas (linajes) en el árbol filogenético. Además, este método tiene la ventaja de escoger, entre un conjunto de modelos, el que mejor se ajuste estadísticamente a los datos, a diferencia de otros métodos donde existe el requisito de especificar *a priori* el modelo. Sin embargo, con la gran cantidad de modelos puede existir un modelo que se ajuste mejor que el modelo seleccionado *a priori*. Para el linaje de Viridiplantae se utilizó el método FEL (Kosakovsky-Pon y Frost, 2005b). El método FEL (Fixed Effects Likelihood) es análogo de métodos que estiman la relación de las tasas de sustitución no sinónimas y sinónimas (dN/dS) para cada uno de los sitios, sin embargo en este método se fijan algunos parámetros como la longitud de rama, lo que permite una disminución de tiempo computacional, pero también puede realizar pruebas de sitio-rama, las cuales se emplean para detectar sitios que están bajo selección en un linaje específico.

5. Resultados

5.1 Reconstrucción filogenética de la familia de factores de inicio de la traducción eIF4E

La reconstrucción filogenética se realizó mediante máxima verosimilitud empleando el alineamiento de Pfam referente a la región “core”, de la familia de factores de inicio de la traducción eIF4E y el modelo de sustitución LG+I+G. El soporte estadístico de *bootstrap* consistió de 100 repeticiones. El árbol obtenido se muestra en la figura 9 con algunas ramas colapsadas para una mejor visualización. Las ramas colapsadas incluyen secuencias pertenecientes a la misma clase estructural, de acuerdo al criterio empleado por Joshi *et al.* (2005) y agrupan secuencias de organismos del mismo grupo taxonómico, al cual le denominamos “linaje orgánico”.

La comparación de la topología del árbol de nuestra reconstrucción filogenética con el árbol obtenido por Joshi *et al.* (2005), observamos que no aparecen los tres clados, referentes a cada una de las clases estructurales, ni los 8 subgrupos que reporta en su trabajo, 4E1 de Metazoa, 4E1 de Fungi, 4E1 de Viridiplantae, 4E2 de Metazoa, nCBP de Viridiplantae, proteínas parecidas a nCBP de Fungi, 4E3 de Metazoa y 4E de Protistas, el clado que incluya a las secuencias denominadas “atípicas” por Joshi.

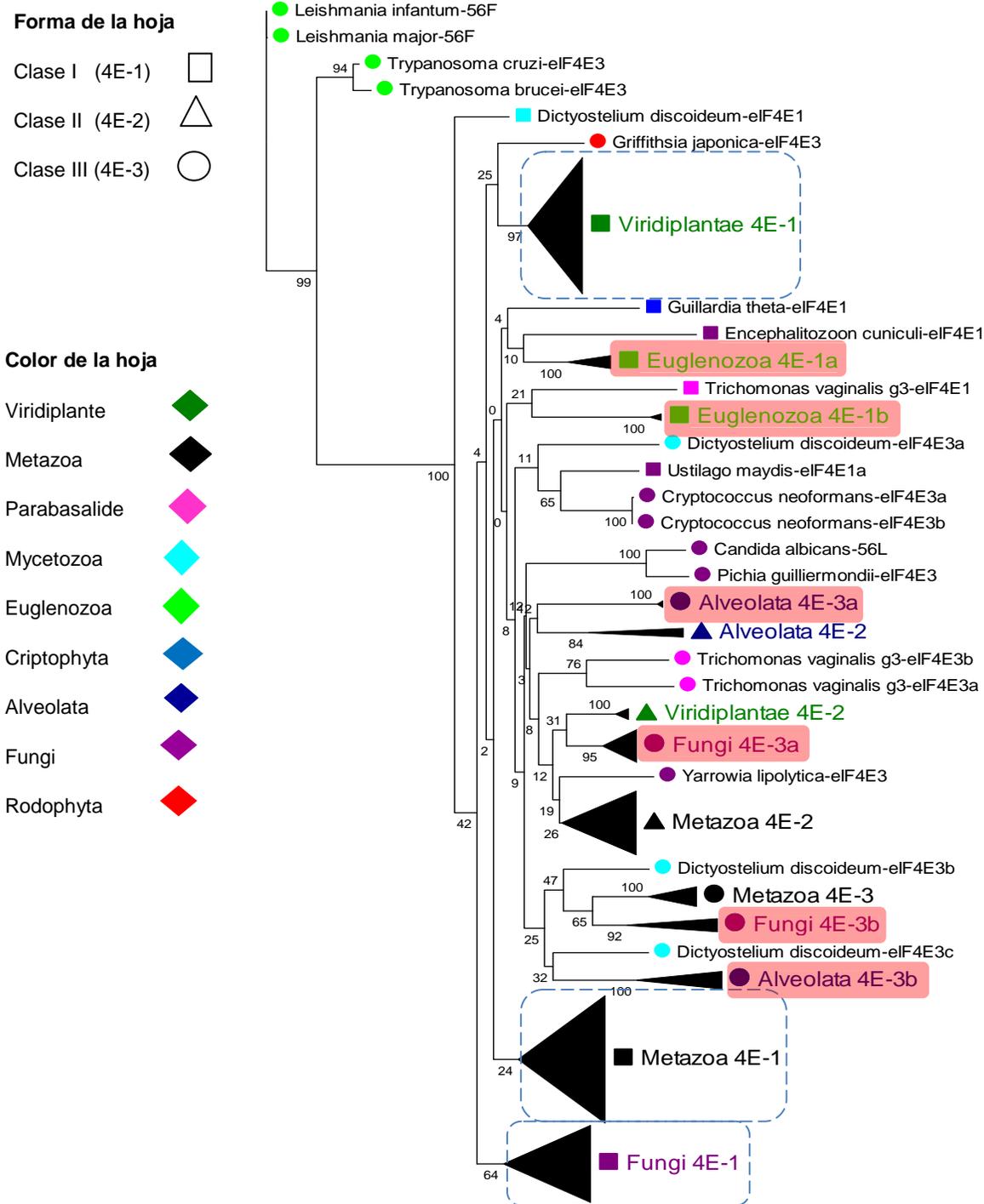


Figura 9. Árbol filogenético sin enraizar construido con el perfil de la región “core” del alineamiento de Pfam, algunas ramas están colapsadas para una mejor visualización. Las ramas colapsadas incluyen secuencias de la misma clase estructural y del mismo linaje orgánico, y están representadas mediante triángulos negros y el tamaño del triángulo es directamente proporcional al número de secuencias que incluye. Las forma de la hoja que indican a que clase estructural pertenece; el cuadrado indica que pertenece a la clase estructural I, el triángulo a la clase II y el círculo a la clase III. El color de la hoja indica el linaje orgánico al cual pertenece. Los números en los nodos internos indican el valor de los *bootstrap*. En los recuadros punteados color azul se indican los 3 clados más grandes de la clase I. Los clados sombreados en rojo indican los clados de la misma clase estructural y del mismo linaje.

Sin embargo, si podemos observar la formación de 3 grandes clados, estos corresponden a las secuencias de la clase estructural I, de los linajes orgánicos de Metazoa, Viridiplantae y Fungi. En este último solamente 2 secuencias se agrupan fuera del clado principal, las cuales corresponden a las especies de *Encephalitozoon cuniculi* y *Ustilago maydis*. El número de secuencias, de la clase estructural I, de estos tres linajes orgánicos es mayor, lo cual ayuda a diferenciar mejor estos clados de los demás y posiblemente la similitud de secuencia de cada uno de los clados es mayor. Con respecto a los demás linajes orgánicos, no se observa la formación de grandes clados debido a que el número de secuencias fue mucho menor, en comparación con los linajes orgánicos anteriores, para la clase estructural I, también se observa la agrupación de secuencias de la clase estructural I en distintos linajes orgánicos y la agrupación de secuencias de la clase estructural I del mismo linaje orgánico en clados distintos, como ocurre con el linaje orgánico de los Euglenozoa (Figura 9, Euglenozoa 4E-1a y Euglenozoa 4E-1b); esto quizás se deba, como lo menciona Joshi *et al.* (2005), a que las secuencias de los protistas presentan inserciones de 12 a 15 residuos de aminoácidos entre los residuos equivalentes a Trp 73 y Trp 102 de eIF4E-1 de *H. sapiens*, y de 4 a 9 entre los residuos equivalentes a Trp 102 y Trp 166 (figura10), lo cual les da una mayor variación de secuencia intralínea por lo que no son agrupados dentro de un sólo clado.

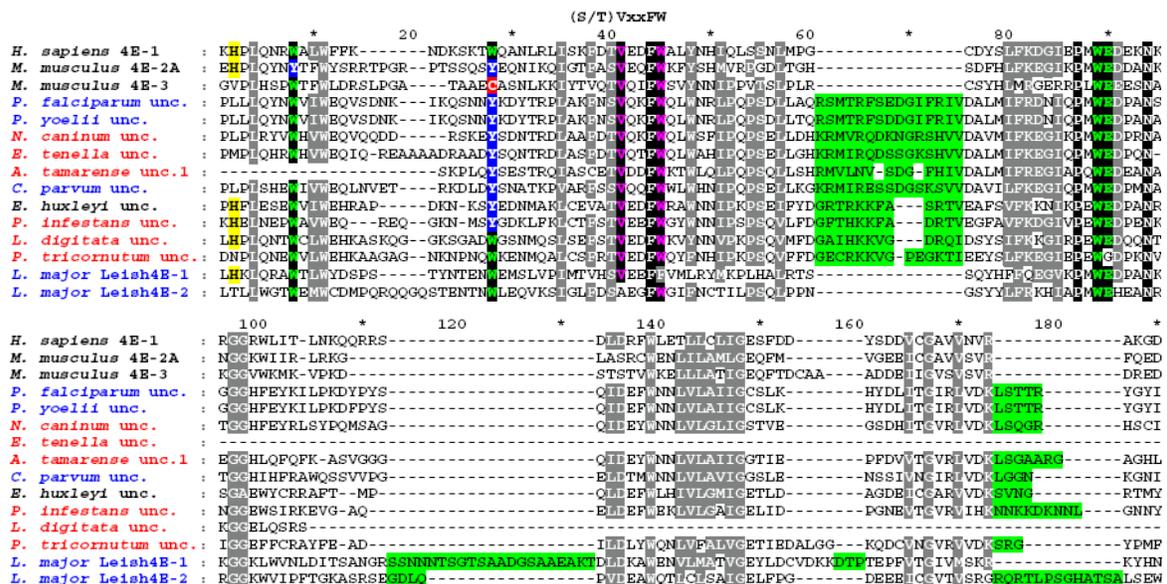


Figura 10. Alineamiento de la región "core" en aminoácidos de secuencias de miembros de linajes orgánicos de Alveolata, Estraptenophiles, Haptophyceae, y de las especies de *Emiliania huxleyi*, *Homo sapiens*, *Mus musculus* 4E-2A y 4E-3. Las inserciones que presentan estas secuencias entre los residuos de Trp 73 y Trp 102 de eIF4E-1 de *Homo sapiens*, y de 4 a 9 entre los residuos equivalentes a Trp 102 y Trp 166 son resaltadas en color verde (Joshi *et al.*, 2005).

Con respecto a la clase estructural II, observamos la formación de clados mejor definidos que incluyen secuencias del mismo linaje orgánico. Con ello asumiríamos que la similitud dentro de cada uno de los linajes orgánicos es mayor, permitiendo esta diferenciación, a pesar de que el número de secuencias en la mayoría de los linajes orgánicos es menor en comparación con la clase estructural I. Hay que subrayar que no se incluyen secuencias de distinto linaje orgánico dentro de algún clado. En esta clase II sobresale el clado de los Metazoa, debido a que es el que contiene mayor número de secuencias, a diferencia de los demás linajes orgánicos que no sobrepasan las 5 secuencias. Por otra parte, observamos que no existen secuencias de la clase estructural II, dentro del linaje orgánico de Fungi, por lo tanto no observamos la formación de un clado que incluya a éstas y aparezca como grupo hermano del clado de Viridiplantae, como se observa en el trabajo de Joshi *et al.* (2005).

Referente a las secuencias de la clase estructural III, encontramos la formación de diversos clados del mismo linaje orgánico y en algunos casos la inclusión de secuencias de distintos linajes orgánicos, exceptuando el clado de las secuencias del linaje orgánico de los Metazoa. Dentro de los linajes orgánicos que presentan más de un clado para las secuencias de la clase estructural III son el Alveolata, y sorprendentemente el de los Fungi, en donde la presencia de secuencias de esta clase estructural III no había sido reportada, de acuerdo al criterio de clasificación estructural y al trabajo de Joshi *et al.* (2005). Tal vez este fenómeno de la formación de diversos clados, se deba a la presencia de una variabilidad considerable de secuencia dentro del linaje, pero con una alta similitud dentro de ciertas secuencias, la cual permite agruparlas dentro de un mismo clado; por otra parte, ahora el linaje que tiene una mayor representatividad, en número de secuencias, es el de los Fungi. Algunos clados de los distintos linajes orgánicos de la clase estructural III pueden ser agrupados dentro de un subárbol, que incluye a los clados de Metazoa, Alveolata y Fungi, sin embargo los demás clados, tanto del linaje orgánico de los Fungi y Alveolata se encuentra en otros subárboles que incluyen a los clados de la clase estructural II, dentro de ellos se encuentra el clado con el mayor número de secuencias de Fungis, el cual aparece como grupo hermano del clado de secuencias de la clase estructural II del linaje de Viridiplantae.

En la figura 11 se muestra una imagen solamente de la topología del árbol sin enraizar con las ramas colapsadas y sin longitud de ramas; a diferencia del presentado en la figura

9 ahora las ramas colapsadas no están indicadas con triángulos negros, solamente con el nombre del linaje seguido de la clase estructural con un tamaño de letra mayor.

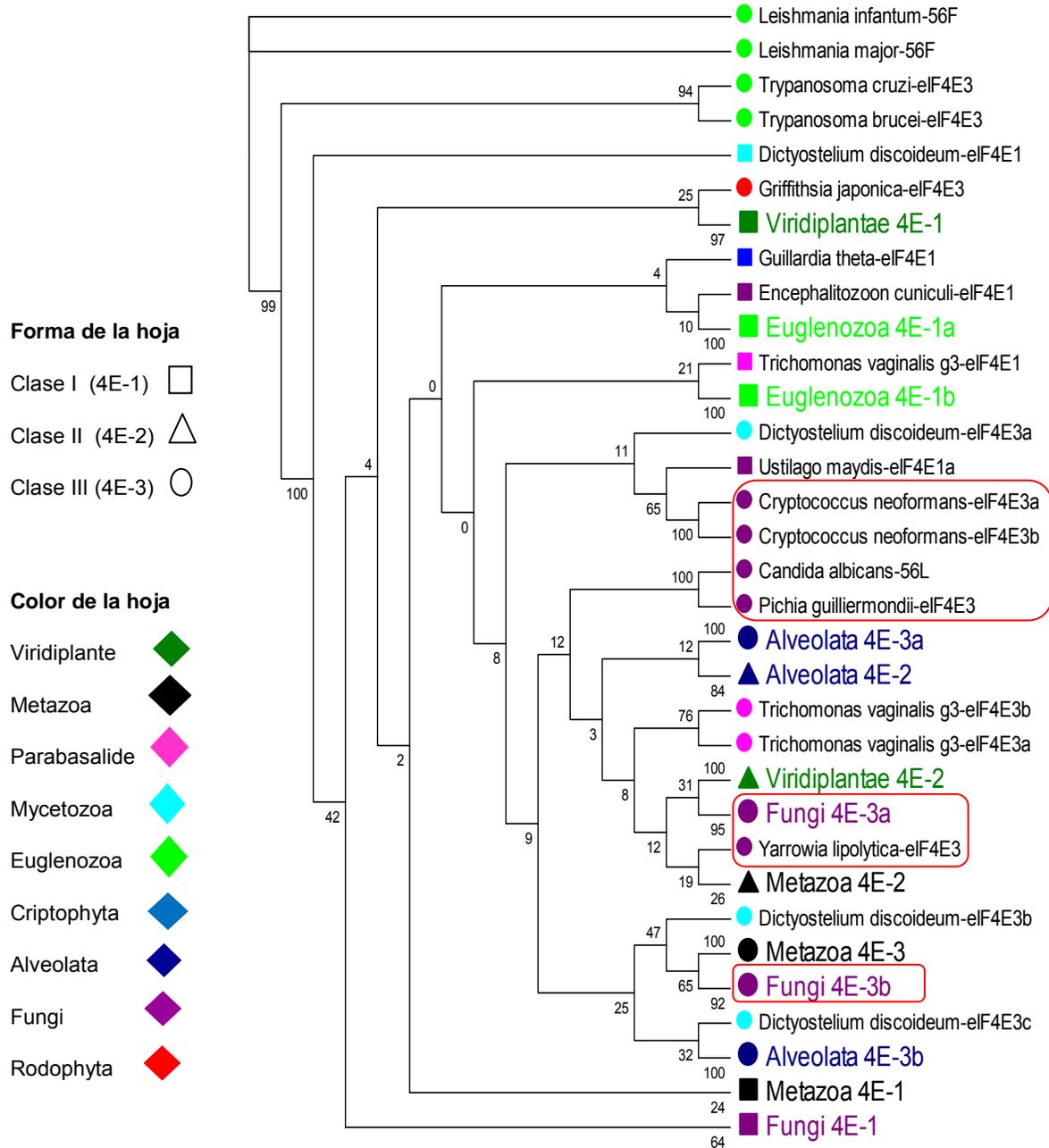


Figura 11. Árbol filogenético sin enraizar, de la región “core” del alineamiento de Pfam con algunas ramas colapsadas. Las ramas colapsadas incluyen varias secuencias de la misma clase estructural y linaje orgánico, y están indicadas por el nombre del linaje orgánico en un tamaño de letra mayor y el color correspondiente al color de la hoja. Los números en los nodos internos indican los valores de *bootstrap*. En recuadro rojo se muestra la clase estructural III del linaje de los Fungi.

Con respecto a los valores de *bootstrap* en algunas ramas son muy bajos, sobre todo en las ramas internas, las cuales llegan a tener hasta valores de cero, lo cual nos indica que esas ramas estadísticamente están pobremente soportadas. Por otra parte encontramos ramas que presentan valores de *bootstrap* de 100 principalmente en las ramas basales de clados formados por secuencias de la misma clase estructural y linaje orgánico.

Al realizar la clasificación se observó que diversas secuencias de los distintos linajes y clases estructurales presentaban sustituciones en algunas posiciones referentes a los 8 triptófanos (43, 46, 56, 73, 102, 113, 130, 166) que caracterizan a los factores eIF4E, las posiciones que presentan sustituciones son 46, 73, 113 y 130, de acuerdo a la numeración de eIF4E1 de *Homo sapiens*; así como en la secuencia consenso que interactúa con 4G y 4EBP, la cual incluye a las posiciones 69, 72 y 73. Las sustituciones que se observaron se muestran en la tabla 2.

Tabla 2 Sustituciones encontradas en las secuencias del alineamiento obtenido de Pfam, en las posiciones 46, 69, 72, 73, 113 y 130, la numeración corresponde a la secuencias de eIF4E1 de *Homo sapiens*

Posición	Aminoácido Esperado	Sustitución
46	Triptófano (W)	Tirosina (Y), Histidina (H), Leucina (L)
69	Valina (V)	Isoleucina(I), Alanina(A), Leucina (L), Cisteína (C)
72	Fenilalanina (F)	Tirosina (Y), Leucina (L), Triptófano (W)
73	Triptófano (W)	Tirosina (Y), Fenilalanina (F), Leucina (L), Alanina (A)
113	Triptófano (W)	Fenilalanina (F), Isoleucina (I), Leucina (L), Valina (V)
130	Triptófano (W)	Alanina (A), Fenilalanina(F)

Los triptófanos de las posiciones 102 y 166 son muy conservados y en sus posiciones homólogas no se presenta alguna sustitución. Esto es lo que se espera ya que el triptófano 102 es clave para la interacción con el *cap*, y el 166 también se ha reportado que interactúa con él, aunque en menor medida (Joshi *et al.*, 2004).

Por otra parte se observó sustituciones distintas a las empleadas para realizar la clasificación estructural, en la posición 56. Estas sustituciones son:

Leucina (L), Valina (V) y Fenilalanina (F)

Estas sustituciones se presentan cuando en la posición 43 se encuentra un residuo de triptófano, tomando en cuenta la clasificación estructural hecha por Joshi *et al.* (2005), podrían ser clasificadas como pertenecientes a la clase estructural III.

De acuerdo a la tabla 2, la mayoría de los residuos de aminoácidos que sustituyen al triptófano original, son residuos no polares, exceptuando a los residuos de histidina y cisteína los cuales son polares. Observamos que en la posición 46, puede haber sustituciones por residuos no polares, tirosina y leucina, y polares como la histidina. Estas sustituciones se presentan principalmente en las clases estructurales I y III, y en distintos linajes orgánicos: Euglenozoa, Viridiplantae, Mycetozoa, Fungi y Metazoa. Por otra parte la incidencia de esta sustitución relativamente es mayor en la clase estructural III, y con mayor frecuencia el residuo de triptófano es sustituido por otro residuo no polar como lo es la tirosina, y por lo contrario la histidina sólo se presenta en una ocasión, lo cual puede ser lógico al ser un residuo polar. No podemos hablar de un patrón de distribución de esta sustitución en los linajes orgánicos, sin embargo la presencia de esta sustitución es mayor en el linaje orgánico de los Fungi.

En la posición 69 el residuo de valina es sustituido por residuos no polares y polares como lo es la cisteína. Las sustituciones se presentan en las 3 clases estructurales aunque con mayor frecuencia en la clase estructural I y III, dentro de estas dos son más comunes en la clase III; en la clase estructural II solamente se presentan en el linaje orgánico de los Alveolata; en los linajes orgánico de Viridiplantae, Cryptophyta y Metazoa las sustituciones se presentan en la secuencias pertenecientes a la clase estructural I mientras que en la clase estructural III la sustituciones se presentan en los linajes orgánicos de Metazoa, Mycetozoa, Parabasalide y con mayor frecuencia en el grupo de los Fungi. En la mayoría de los casos la alanina y la isoleucina sustituyen a la valina, ambos residuos no polares. El residuo alanina es más común encontrarlo en

secuencias de Euglenozoa, Alveolata y Fungi, mientras que la isoleucina en los linajes orgánicos de Viridiplantae, Metazoa y Fungi. La cisteína es la sustitución que sólo se presenta en dos ocasiones, una en el grupo de los Parabasalides y en el de los Fungi.

Con respecto a la fenilalanina en la posición 72, es sustituida solamente por residuos no polares. Las sustituciones se presentan en las 3 clases estructurales aunque con menor presencia en la clase I, ya que se encuentra solo en una secuencia de Metazoa. La distribución de las sustituciones en las otras clases parece tener un patrón o ser exclusivas de ciertos grupos, en la clase estructural II la sustitución de la fenilalanina por triptófano solamente se presenta en el grupo de los insectos dentro de los Metazoa, mientras que la sustitución por leucina sólo ocurre en secuencias de la clase estructural III en los Fungi, aunque en clados distintos.

Por otra parte, el triptófano en la posición 73, que está dentro de los 8 residuos de triptófano característicos del factor de inicio de la traducción eIF4E y que junto con los residuos de valina y fenilalanina en las posiciones 69 y 72, forma parte de la secuencia consenso que interactúa con 4G y 4EBP's, y puede sustituirse por residuos no polares. Las sustituciones solamente se presentan en miembros de las clases estructurales I y III, con un poco de mayor frecuencia en la III; las sustituciones se encuentran en las secuencias de los linajes orgánicos de los Metazoa, los Euglenozoa, los Cryptophyta, los Alveolata y los Fungi, y en este último es más común encontrar esta sustitución. Algunas sustituciones son características de ciertos clados, por ejemplo la fenilalanina es la sustitución más frecuente y se observa en un clado de los Euglenozoa, miembros de la clase estructural I y en un clado de los Fungi que incluye secuencias de la clase III. La alanina es la segunda sustitución con mayor frecuencia y es característica de las secuencias de la clase estructural III de los Alveolata.

El residuo de triptófano en la posición 113 puede sustituirse por residuos no polares, y se presenta en las 3 clases estructurales, pero más frecuente en la clase III. Los linajes orgánicos en los cuales se puede encontrar secuencias con esta sustitución son los Euglenozoa, los Alveolata, los Parabasalide, los Mycetozoa y los Fungi; en este último solamente se encuentra en una secuencia, perteneciente a la clase estructural I. Las sustituciones más frecuentes son la fenilalanina y la isoleucina, en algunos grupos estas sustituciones son características como en el clado los Alveolata de la clase estructural III.

Referente al triptófano de la posición 130, solamente 2 residuos de aminoácidos se observaron como sustituciones en esta posición, ambos no polares. Las sustituciones se presentaron en secuencias pertenecientes a la clase estructural III y en 3 linajes: los Euglenozoa, los Fungi y los Mycetozoa. La sustitución más frecuente es por fenilalanina, y ocurre en los Euglenozoa y en algunos Fungi. La sustitución del triptófano 130 por alanina se encontró en una secuencia de los Mycetozoa.

Debido a lo anterior se elaboraron dos filogenias, en las cuales se eliminaron del alineamiento aquellas secuencias que presentaban sustituciones, con la finalidad de obtener un árbol más resuelto. La metodología empleada para la obtención de estos árboles fue la misma que se describió en la figura 5.

En la primera de estas dos filogenias, se eliminaron del alineamiento las secuencias que presentaban sustituciones en las posiciones 46, 113 y 130, y solamente se permitió variabilidad en la secuencia consenso de unión a 4G y 4EBP, posiciones 69, 72 y 73. El modelo de sustitución que se utilizó fue el JTT+I+G y el soporte estadístico se evaluó con *bootstrap* de 100 repeticiones. El árbol que se obtuvo en esta segunda reconstrucción se presenta en la figura 12, algunas ramas están colapsadas para una mejor visualización. Las ramas colapsadas agrupan las secuencias de la misma clase estructural y del mismo linaje, y están representadas con un triángulo negro, cuyo tamaño es directamente proporcional al número de secuencias que agrupa.

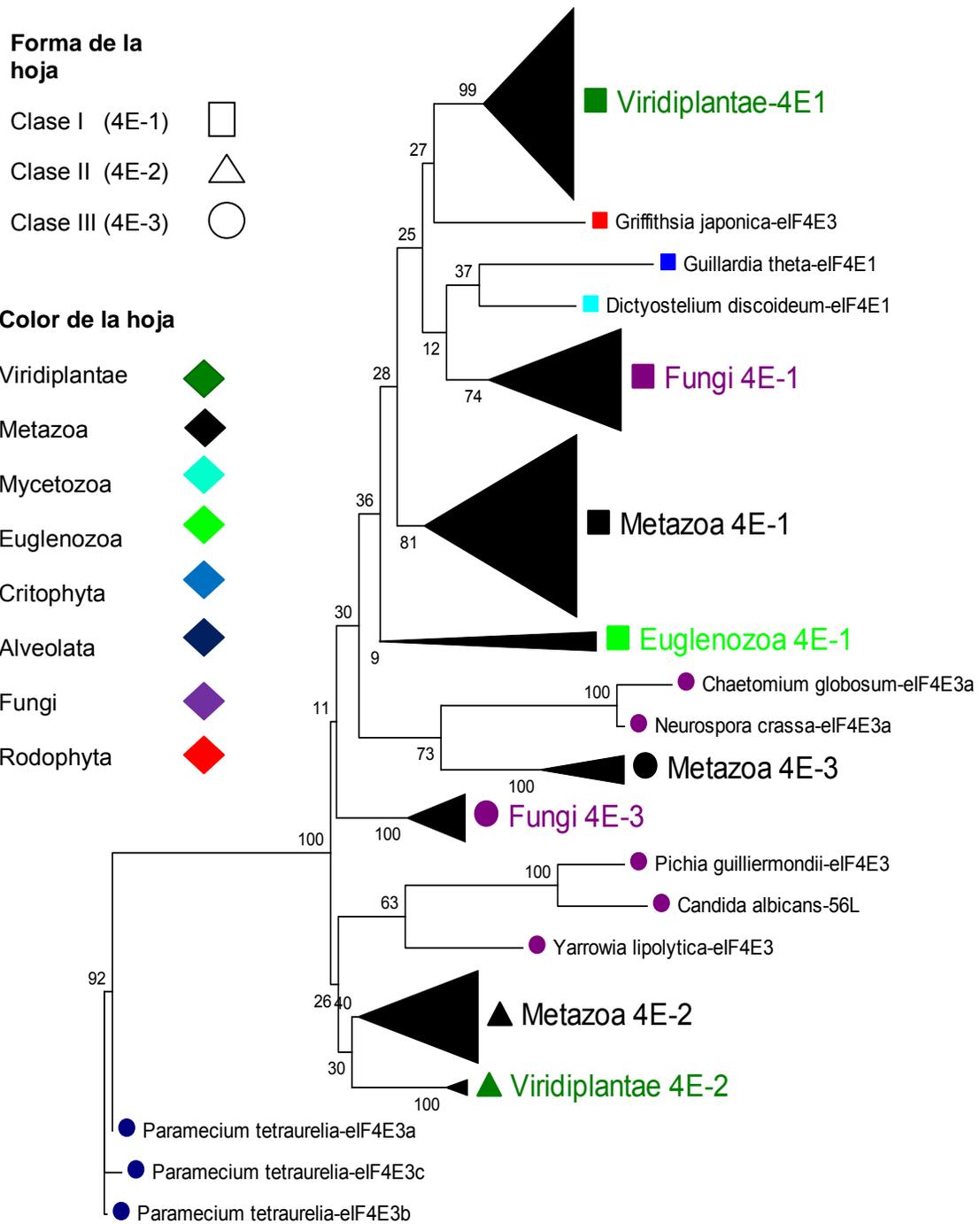


Figura 12. Filogenia sin enraizar con algunas ramas colapsadas obtenida con el alineamiento de las secuencias que sólo presentan sustituciones en la secuencia consenso de unión a 4G y 4EBP, posiciones 69,72 y 73. Las ramas colapsadas están representadas con triángulos negros y el tamaño del triángulo es directamente proporcional al número de secuencias que agrupa. Las ramas colapsadas agrupan secuencias de la misma clase estructural y linaje orgánico. La forma de la hoja que indican a que clase estructural pertenece; el cuadrado indica que pertenece a la clase estructural I, el triángulo a la clase II y el círculo a la clase III. El color de la hoja indica el linaje orgánico al cual pertenece. Los números en los nodos internos indican el valor de de soporte de bootstrap.

En el árbol anterior observamos la formación de clados bien diferenciados, que incluyen secuencias de la misma clase estructural y del mismo linaje, dentro de ellos sobresalen por su tamaño, el número de secuencias que agrupan, los clados referentes a la clase estructural I de los linajes de Viridiplantae, Fungi y Metazoa. Así mismo, podemos notar que todos los clados formados por las secuencias de la clase estructural I de los diversos linajes orgánicos (Viridiplantae, Fungi, Metazoa y Euglenozoa) se localizan en un subárbol (Figura 13 subárbol A) y que tiene como grupo hermano al subárbol formado por el clado de la clase estructural III de Metazoa y una de las dos secuencias que presentan *Chaetomium globosum* y *Neurospora crassa* de la clase estructural III, a las cuales les hemos denominado eIF4E3a (Figura 13 subárbol B). Lamentablemente el soporte estadístico para esta relación filogenética es muy bajo. El clado que incluye la mayoría de las secuencias de la clase estructural III del linaje de los Fungi aparece como grupo hermano de los dos subárboles anteriores (subárbol A y B), y en él encontramos la segunda secuencia de *Chaetomium globosum* y *Neurospora crassa* eIF4E3b. Con respecto a las secuencias de la clase estructural II, están agrupadas dentro de dos clados que corresponden a los linajes orgánicos de Viridiplantae y Metazoa, y al igual que las secuencias de la clase I, se encuentran dentro del mismo subárbol (figura 13 subárbol C) y que tiene como grupo hermano al subárbol formado por las secuencias de la clase estructural III de *Pichia guillermondii*, *Yarrowia lipolytica* y *Candida albicans*. En la secuencia de esta última, en la posición correspondiente al triptófano 56, de acuerdo a la numeración de eIF4E de *H. sapiens*, presenta un residuo de leucina, que es un residuo distinto a los característicos de las secuencias de la clase estructural III (cisteína o tirosina) de acuerdo a Joshi *et al* (2005), pero, a pesar de ello, se agrupa en la filogenia con secuencias de la clase estructural III del mismo linaje orgánico (Fungi), con lo cual podemos inferir que podría incluirse dentro de la clase estructural III. Las únicas secuencias que pertenecen a la misma clase estructural y al mismo organismo que no forman un clado, son las secuencias de *Paramecium tetraurelia* a las cuales les hemos denominado eIF4Ea, eIF4Eb y eIF4Ec para poder diferenciarlas, sin embargo esto podría deberse a que el árbol no está enraizado.

Para una mejor visualización de lo anterior en la figura 13 se muestra una imagen, de la topología del árbol sin enraizar y sin longitudes de ramas, pero a diferencia del presentado en la figura 12, ahora las ramas colapsadas no se indican con triángulos negros, sino con el nombre del linaje seguido de la clase estructural con un tamaño de letra mayor.

También podemos notar que los valores de *bootstrap* para las ramas más profundas aumentaron en comparación con la primera filogenia, ya que no se presentan ramas con valores de cero, y al igual que en la primera filogenia se observan valores de bootstraps de 100 principalmente en las ramas basales de clados formados por secuencias de la misma clase estructural y del mismo linaje. Al parecer la eliminación de secuencias con sustituciones en las posiciones 46, 113, 130 permite obtener un árbol con clados más definidos, reduciendo la agrupación de secuencias que pertenecen a clases y linajes orgánicos distintos, así como ramas con mejor soporte estadístico.

Con respecto a la segunda filogenia en ella, no se permitió ninguna sustitución en las secuencias, más que aquellas referentes a las posiciones 43 y 56, por lo tanto sólo debían de presentar triptófanos en las posiciones 46, 73, 102, 113, 130 y 166, valina en la posición 69 y fenilalanina en la posición 72, residuos característicos de la secuencia consenso. La metodología empleada para la reconstrucción de esta filogenia fue la misma que se muestra en la figura 1, con la variante que el alineamiento empleado como segundo perfil, es el alineamiento de secuencias carentes de sustituciones en las posiciones antes mencionadas. El modelo de sustitución empleado fue JTT+I+G y el soporte estadístico se evaluó con *bootstrap* de 100 repeticiones. La filogenia obtenida sin enraizar se muestra en la figura 14, algunas ramas están colapsadas para una mejor visualización. Las ramas colapsadas agrupan secuencias de la misma clase estructural y del mismo linaje orgánico, y están representadas con triángulos negros, en donde el tamaño de los triángulos es directamente proporcional al número de secuencias que agrupa.

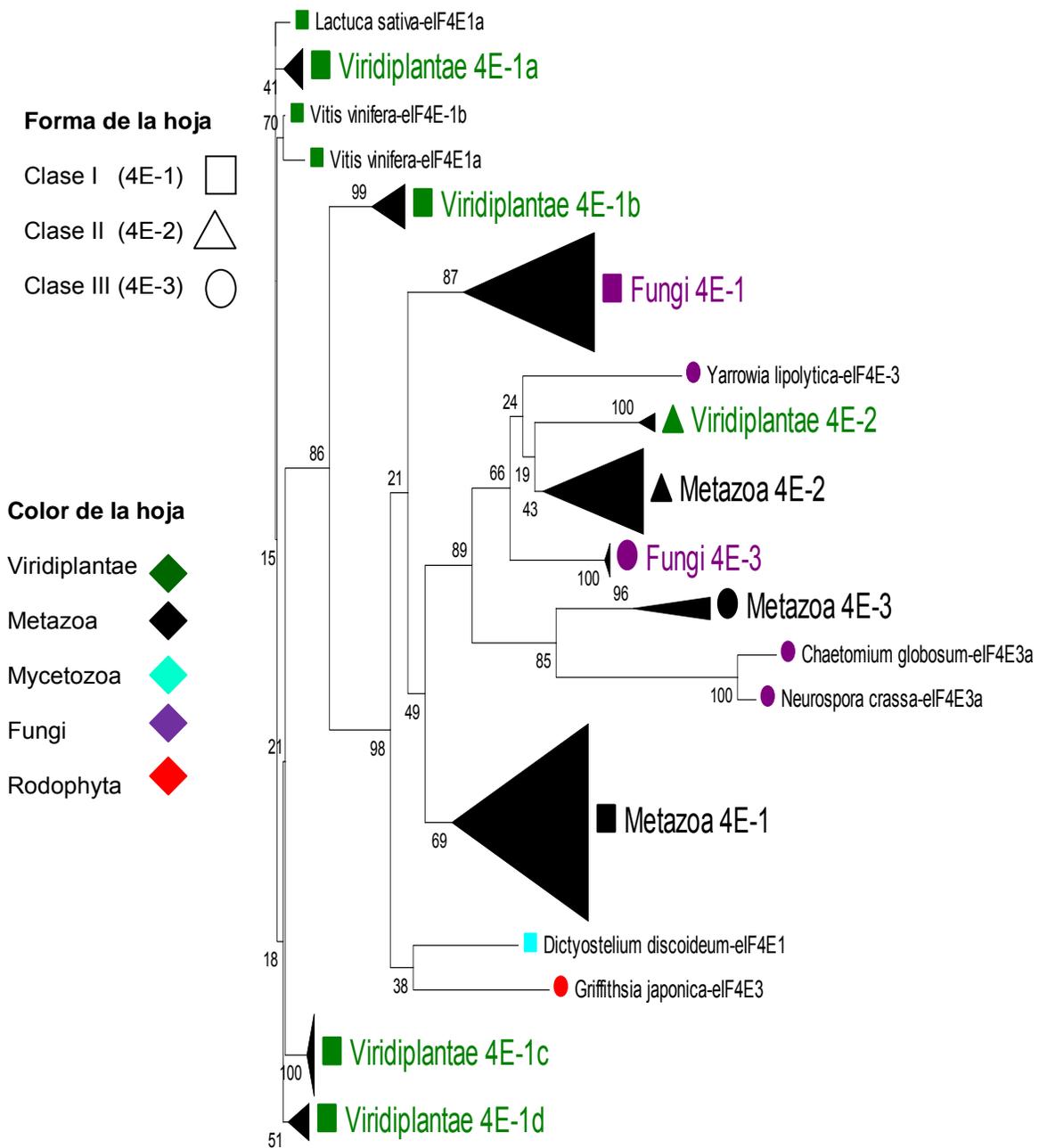


Figura14. Árbol filogenético sin enraizar construido con la región core del alineamiento de secuencias sin sustituciones en las posiciones 46, 69, 72, 73, 113 y 130; algunas ramas están colapsadas para una mejor visualización. Las ramas colapsadas incluyen secuencias de la misma clase estructural y del mismo linaje orgánico, y están representadas mediante triángulos negros y el tamaño del triángulo es directamente proporcional al número de secuencias que incluye. La forma de la hoja indica a que clase estructural pertenece y el color de la hoja indica el linaje orgánico. Los números en los nodos internos indican el valor de *bootstrap*.

En la figura anterior se presenta esta última reconstrucción, y en donde podemos observar que ciertas características que encontramos en la reconstrucción filogenética que se hizo con la región *core* del alineamiento de secuencias, que presentaban sustituciones en las posiciones homólogas a las posiciones 69, 72 y 73 de eIF4E de *H. sapiens*, destacan por su tamaño los clados de la clase estructural I de Fungi, Metazoa y de la clase estructural II el clado de Metazoa. Al igual que en la filogenia pasada (figura 12), se encuentra como grupo hermano del clado de secuencias de la clase estructural II del linaje de Viridiplantae, por lo tanto todas las secuencias de la clase estructural II se encuentran en un subárbol (figura 15 subárbol A), que a su vez tiene como grupo hermano, a diferencia de la reconstrucción pasada, una sola secuencia de la clase estructural III de Fungi, esta secuencia pertenece a *Y. lipolytica*. Otra diferencia, es que el subárbol formado por las secuencias de la clase estructural II y la secuencia de la clase estructural III de *Y. lipolytica* tienen como grupo hermano el clado de las secuencias de la clase III de Fungi, sin embargo, el número de secuencias que incluye es menor en comparación con el árbol anterior. Por otra parte, este clado aparecía como grupo hermano del subárbol que agrupaba todas las secuencias de la clase estructural I y de la clase estructural III de Metazoa y Fungi en la reconstrucción de la figura 12. Un subárbol que permanece constante es el formado por el clado de la clase estructural III de Metazoa con las secuencias de eIF4E3a de *Chaetomium globosum* y *Neurospora crassa* (figura 15 subárbol B). Una diferencia notoria, en comparación con la filogenia anterior (figura 13), es que las secuencias de la clase estructural I no se encuentran en un sólo subárbol y las secuencias de la clase estructural I del linaje de Viridiplantae se encuentran en varios clados dispersos en el árbol. Sin embargo, esto puede deberse a que el árbol no se encuentra enraizado, ya que si la raíz estuviera entre la rama que agrupa a las clases estructurales II y III por un lado y la rama que agrupa a la clase estructural I por otro, las secuencias de la clase I de Viridiplantae se agruparían formando un solo clado. Al parecer el mayor número de secuencias que presentaban sustituciones en las posiciones homólogas a valina 69, triptófano 72 y fenilalanina 73 de eIF4E de *H. sapiens* pertenecía al clado de la clase estructural III de los linajes orgánicos de Fungi y Metazoa, ya que su número en esta última filogenia disminuyó con respecto a reconstrucción realizada con el alineamiento de secuencias que presentaban sustituciones en las posiciones anteriormente mencionadas. En la figura 15 se muestra una imagen de la topología del árbol sin enraizar, con sus valores de *bootstrap* y sin longitudes de rama, para una mejor visualización de los clados. Sin embargo, las ramas colapsadas no se indicaban con

triángulos negros y solamente se etiquetan con el nombre del linaje seguido de la clase estructural con un tamaño de letra mayor.

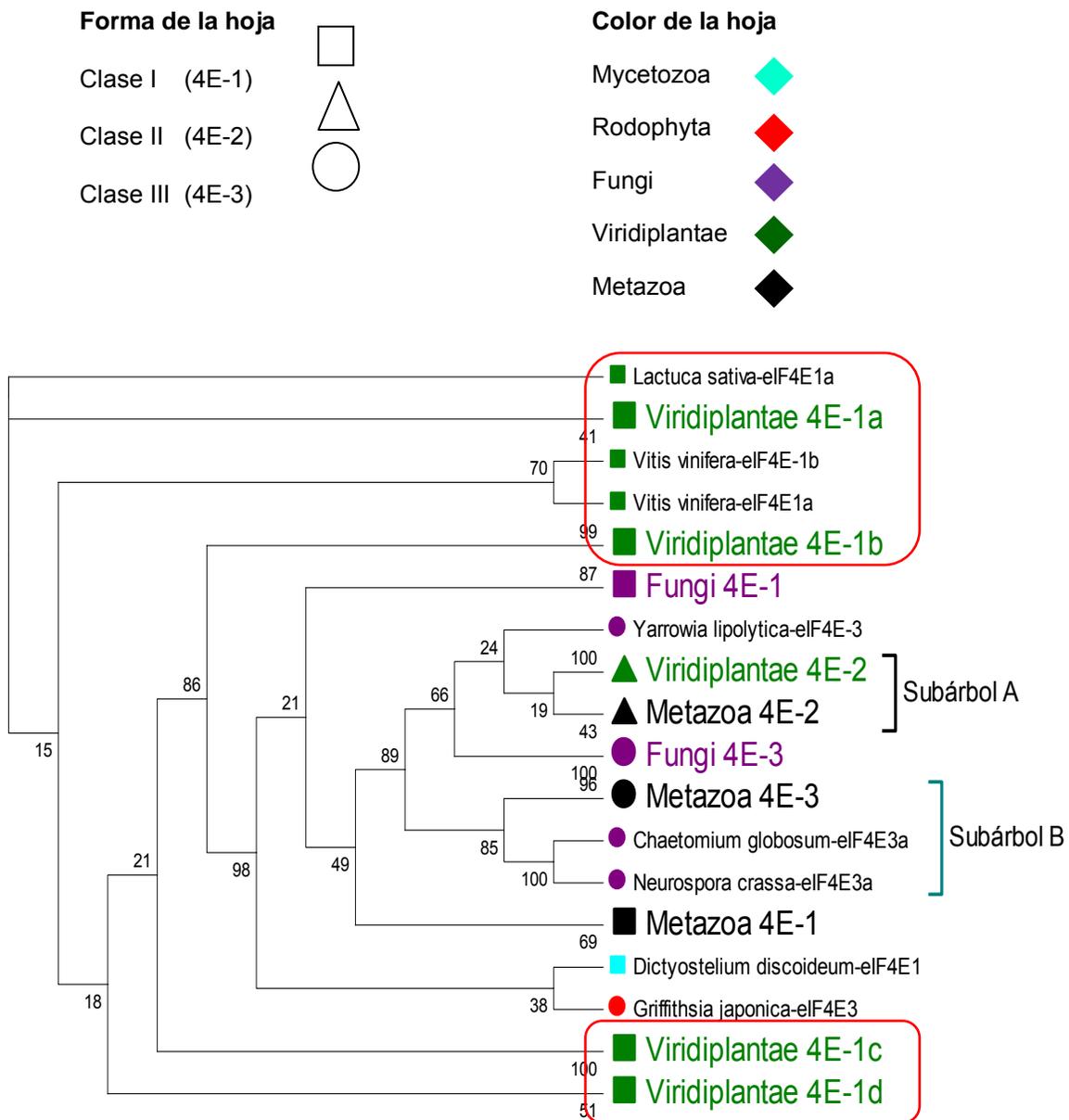


Figura 15. Filogenia sin enraizar con algunas ramas colapsadas obtenida con la región *core* del alineamiento de secuencias sin sustituciones en las posiciones 46, 69, 72, 73, 113 y 130. Las ramas colapsadas incluyen varias secuencias de la misma clase estructural y linaje orgánico, y están indicadas por el nombre del linaje seguido de la clase estructural al que pertenecen con un tamaño de letra mayor. La forma y color de la hoja indican a la clase estructural y al linaje orgánico al que pertenecen. Los números en los nodos internos indican los valores de *bootstrap*. El subárbol A agrupa a las secuencias de la clase estructural II y el subárbol B agrupa al clado de la clase estructural III de Metazoa y 2 secuencias de Fungi. Los recuadros en rojos indican los diversos clados de la clase estructural I del linaje orgánico de Viridiplantae

Podemos observar que los valores de bootstrap aumentaron significativamente en algunas ramas profundas en comparación a la filogenia anterior. Al igual que en el árbol de la figura 12 se observan valores de bootstraps de 100 principalmente en las ramas basales de clados formados por secuencias de la misma clase estructural y del mismo linaje. Aunque los soportes de todas las ramas no son muy altos podrían considerarse mejor soportados estadísticamente en comparación con las 2 reconstrucciones anteriores (figura 9 y 12). Al eliminar las secuencias que presentaban sustituciones, la cantidad de linajes incluidos en la reconstrucción de la filogenia disminuyen sesgándose hacia secuencias pertenecientes a los linajes de Metazoa, Viridiplantae y Fungi; sólo una secuencia del linaje de Mycetozoa y una de Rodophyta se mantiene en esta última filogenia, las cuales se agrupan en un clado.

5.2 Enraizamiento del Árbol Filogenético

Los métodos de reconstrucciones filogenéticas nos dan como resultado árboles no enraizados, es decir solamente diagramas de ramificación a partir de los cuales es difícil establecer el orden de los eventos de cambio y las relaciones de ancestría-descendencia. El enraizamiento de un árbol filogenético nos permite darle una polaridad o dirección e inferir el orden de estos eventos y relaciones. Un árbol enraizado tiene un nodo identificado como la raíz desde el cual todos los demás nodos descienden, por ello un árbol enraizado tiene dirección.

Mediante el programa Notung (Vernot, *et al* 2005) se enraizaron los árboles obtenidos de las tres reconstrucciones (figuras 9, 12 y 14). Este programa emplea el método de árboles reconciliados para llevar a cabo el enraizamiento y permite elegir de una serie de ramas aquella que será la raíz, de acuerdo a la puntuación del programa. El árbol enraizado de la primera reconstrucción se muestra en la figura 16 mediante una representación de círculo.

Observamos que al enraizar el árbol se revelan nuevas relaciones filogenéticas que no se observaban en la reconstrucción sin enraizar, probablemente se deba a que el árbol no tenía una dirección. Podemos distinguir claramente la formación de dos ramas principales, la primera que tiene el mayor número de secuencias, agrupa a todas las secuencias de la clase estructural I y 5 secuencias de la clase estructural III; dentro de esta notamos la formación de 3 clados bien definidos, que incluyen solo secuencias del mismo linaje orgánico, y que corresponden a los linajes de los Metazoa, los Viridiplantae y los Fungi. En este último linaje solamente una secuencia no se agrupa dentro de este clado. Respecto a las relaciones filogenéticas notamos la formación de un subárbol (al que denominaremos **subárbol A**), que incluye al clado de los Fungi y como grupo hermano 4 secuencias pertenecientes a la clase estructural III del linaje de los Euglenozoa que corresponden a *Trypanosoma brucei*, *Trypanosoma cruzi*, *Leishmania infantum* y *Leishmania major* (las secuencias de estos dos últimos organismos presentan en la posición homóloga a la posición 56 de eIF4E de *H. sapiens* un residuo de fenilalanina) y una secuencia de la clase estructural I del linaje de Micetozoa; este subárbol A formado por el clado de los Fungi y el clado de los Euglenozoa y Micetozoa se observa como grupo hermano del subárbol (al que denominaremos **subárbol B**) que incluye el clado del linaje orgánico de Viridiplantae y una secuencia de la clase estructural III que corresponde a *Griffithsia japonica*, algo que no se notaba en el árbol sin enraizar. Por otra parte, el clado del linaje orgánico de Metazoa se encuentra como grupo hermano del subárbol (al que denominaremos **subárbol C**) formado por el subárbol A y el subárbol B. Otro subárbol que observamos como grupo hermano del subárbol formado por el linaje de Metazoa y el subárbol C, es aquel que incluye las secuencias de los linajes orgánicos de Euglenozoa, Criptophyta y la secuencia de *Encephalitozoon cuniculi* que pertenece a Fungi, y por último observamos la formación de otro subárbol que agrupa otras secuencias del linaje de Euglenozoa y una de Parabasalide. En los subárboles que incluyen secuencias del linaje de Euglenozoa, encontramos organismos que presentan más de una secuencia de la clase estructural I y que se agrupan en clados distintos, como es el caso de *Leishmania infantum* y *Leishmania braziliensis*. A las secuencias de estos organismos se les denominó eIF4Ea y eIF4Eb respectivamente. En el subárbol que incluye las secuencias de los linajes orgánicos de los Euglenozoa, los Criptophyta y una de los Fungi, se encuentran las secuencias eIF4Eb de cada organismo, por otra parte las secuencias de eIF4a de *Leishmania infantum* y *Leishmania braziliensis* se agrupan en el subárbol formado por secuencias del linaje orgánico de Euglenozoa y de

Parabasalide, posiblemente esto se deba a que las secuencias (eIF4a y eIF4b) presenten una alta variación a nivel de secuencia y por lo tanto son agrupadas en clados distintos, en donde la similitud de secuencia dentro del clado es mayor (figura 17).

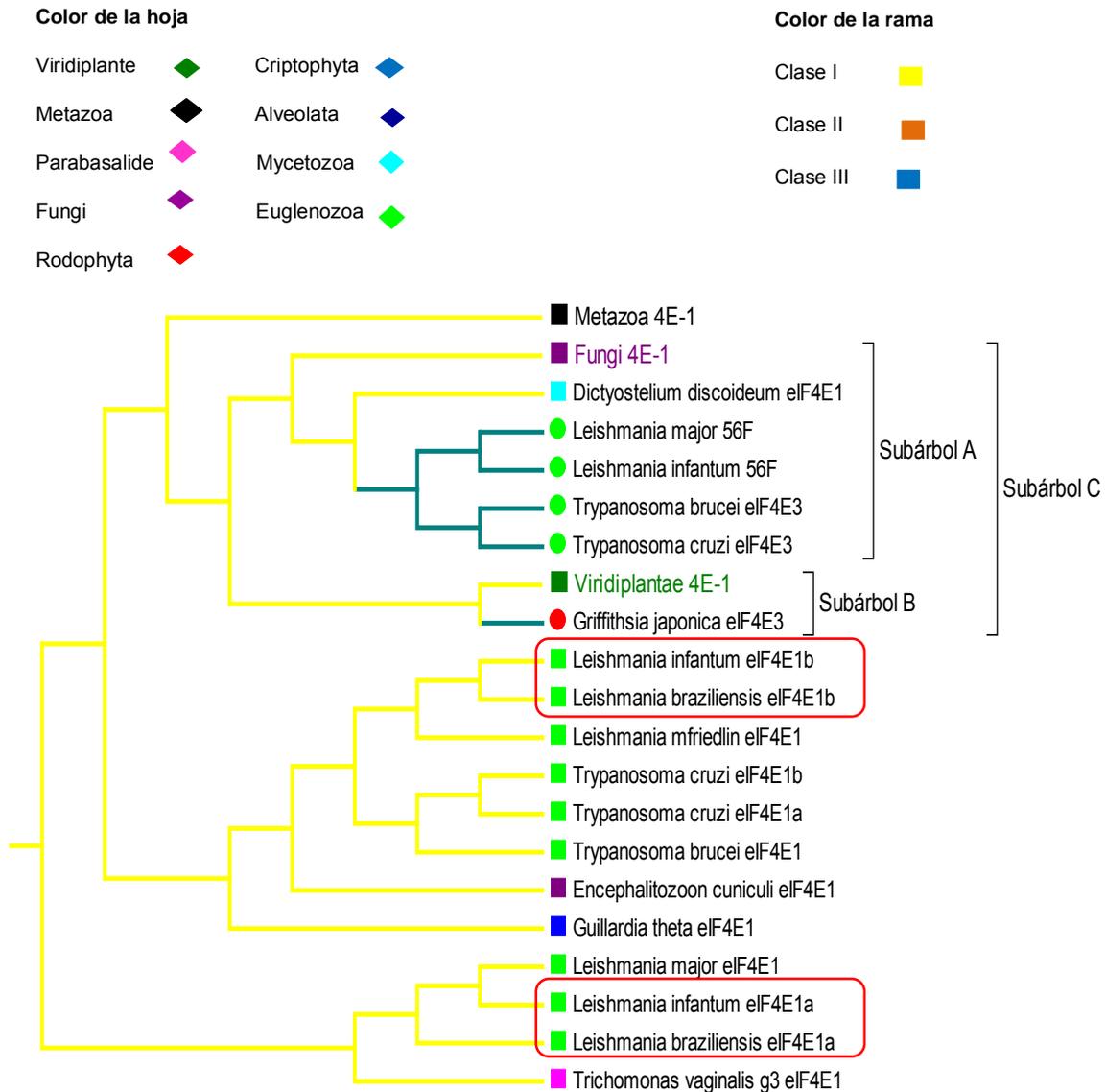


Figura 17. Topología de la rama que incluye a todas las secuencias de la clase estructural I, del árbol enraizado presentado en la figura 16, con las ramas colapsadas. Las ramas colapsadas agrupan secuencias del mismo linaje, y son indicadas con el nombre del linaje seguido de la clase estructural. El color de la rama indica la clase a la que pertenecen. La forma y el color de la hoja indica la clase y el linaje al que pertenecen. Los corchetes indican los subárboles, el Subárbol A incluye el clado de Fungi, una secuencia del linaje Mycetozoa y 4 secuencias de la clase estructural III de Euglenozoa, el Subárbol B incluye el clado de Viridiplantae y una secuencia de la clase estructural III del linaje organismo de Rodophyta, el Subárbol C incluye los Subárboles A y B. En los recuadros de color rojo se indican las secuencias eIF4Ea y eIF4Eb de *Leishmania infantum* y *Leishmania braziliensis*.

La segunda rama que observamos al enraizar el árbol presentado en la figura 16 incluye las secuencias pertenecientes a la clase estructural II y III, a diferencia de la primera rama, encontramos que la distribución de los clados que se observaban en el árbol sin enraizar (figura 9) aún se mantiene. Continuamos observando la formación de varios clados del mismo linaje orgánico y clase estructural, como el caso del linaje de Fungi o la formación de clados que incluyen secuencias de diferentes clase estructural y del mismo linaje orgánico, como el que incluye secuencias de la clase estructural II y III del linaje orgánico de los Alveolata; así mismo observamos con mayor claridad la formación de un subárbol que incluye solamente secuencias de la clase estructural III de los linajes orgánicos de Metazoa, Fungi, Mycetozoa y Alveolata, y la formación del subárbol que incluye las secuencias de la clase estructural II, exceptuando las del linaje orgánico de Alveolata. En este subárbol encontramos la formación de dos clados, uno que agrupa las secuencias de la clase estructural II de Metazoa y como grupo hermano la secuencia de la clase estructural III de *Y. lipolytica*, y el segundo que agrupa a las secuencias de la clase estructural II de Viridiplantae y un clado de secuencias de la clase estructural III de Fungi.

Con respecto a los árboles enraizados de las dos reconstrucciones restantes, las relaciones filogenéticas que se observan en los árboles sin enraizar, se hacen más visibles: en algunos casos se agrupan los diversos clados de la misma clase estructural y linaje dentro de un solo clado, como se observa en la última filogenia con las secuencias de la clase estructural I del linaje de Viridiplantae o con el linaje de Alveolata en la segunda reconstrucción. El árbol enraizado de la reconstrucción filogenética que se realizó con el alineamiento de secuencias de la región *core* que presentan sustituciones en las posiciones 69, 72 y 73 se muestra en la figura 18, mientras que el árbol enraizado del árbol obtenido de la reconstrucción filogenética hecha con el alineamiento de secuencias de la región *core* sin sustituciones en las posiciones 46, 69, 72, 73, 113 y 130 se muestra en la figura 19.

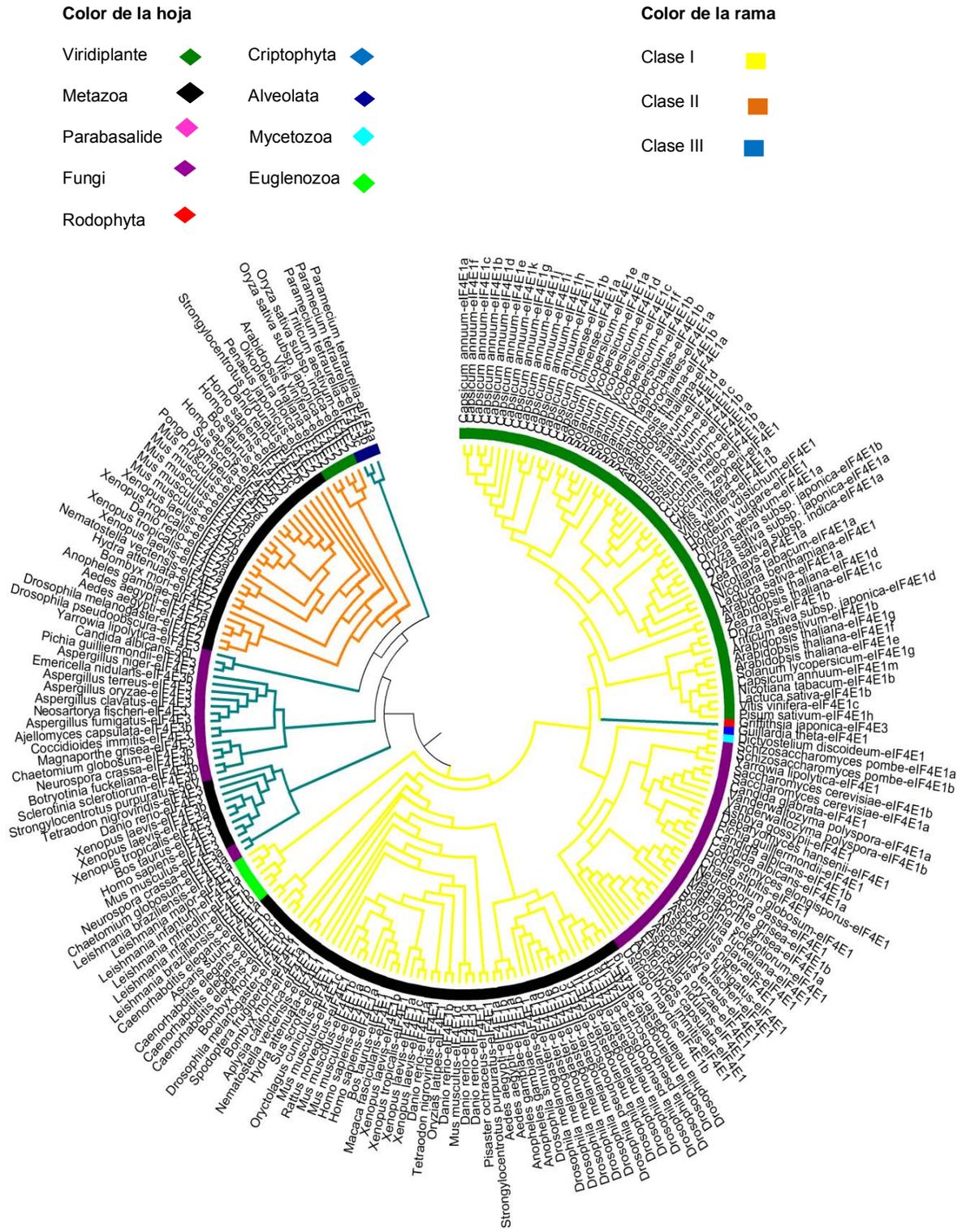


Figura 18. Árbol enraizado representado en forma circular de la reconstrucción filogenética empujando el alineamiento de secuencias de la región core con sustituciones en las posiciones 69, 72, 73, mediante el método de arboles reconciliados; el color de la hoja indica el linaje organizmico al que pertenece y el color de la rama a que clase estructural pertenecen.

De acuerdo a los tres árboles enraizados de las reconstrucciones filogenéticas, encontramos ciertos patrones que permanecen constantes en la topología. Dentro de éstos observamos la formación de dos ramas principales, una que agrupa a las secuencias de la clase estructural I, la cual es más evidente en los dos últimos árboles enraizados (figuras 18 y 19), ya que sólo se presenta la inclusión de una secuencia de la clase estructural III de *G. japonica*. Así mismo observamos que en esta rama los clados formados por las secuencias de los linajes de los Viridiplantae y de los Fungi se encuentran como grupos hermanos formando un subárbol, mientras que el clado de los Metazoa lo encontramos como grupo hermano del subárbol, que incluye al clado de Viridiplantae y Fungi. Con respecto a la segunda rama notamos que agrupa a los miembros de la clase estructural II y III, y en los tres árboles enraizados observamos la formación de un subárbol que agrupa a las secuencias de la clase estructural III de Metazoa y Fungi. Por otra parte la formación de un subárbol que incluye a las secuencias de la clase estructural II se observa más claramente en los dos últimos árboles enraizados (figuras 18 y 19), y este subárbol (compuesto por los clados de Metazoa y Viridiplantae) siempre aparece como grupo hermano de un clado formado de secuencias de la clase estructural III del linaje de Fungi.

Con base en la formación de las dos ramas principales que se observan en los tres árboles enraizados, se realizó una búsqueda para encontrar que posición y residuos son característicos para cada una de las ramas, mediante el empleo del programa Mesquite (Maddison y Maddison, 2007). Con este programa mapeamos que residuo se presentaba en cada una de las posición del alineamiento de secuencias de la región denominada *core* a través del árbol enraizado de la primera reconstrucción filogenética. El alineamiento presenta una gran variabilidad en las secuencias, pero a pesar de ello identificamos algunas posiciones que presentan sustituciones características que podrían ser tomadas en cuenta para diferenciar ambas ramas, estas posiciones y sustituciones se muestran en la tabla 3.

Tabla 3. Sustituciones en las posiciones que pueden ser empleadas para diferenciar las ramas principales, la numeración es de acuerdo a la secuencia de eIF4E de *H. sapiens*. La rama denominada "1" corresponde a la rama que incluye a todas las secuencias de la clase estructural I y 5 secuencias de la clase estructural III (4 del linaje de Euglenozoa y una de Rodophyta), mientras que la rama denominada "2" incluye a todas las secuencias de la clase estructural II y III

Posición en la secuencia	Rama 1	Rama 2
56	W Y C	YCWLFLV
116	STQPMNVIALKCHGFD	RKHLTNS
162	KRT	IVLRQSHM
167	TVLG	NTHV FYLASI
180	GAERLKS	RGMTSLVEP
198	IQSFETNMVYHDP	KRQPSVIYN

De acuerdo a la tabla anterior notamos que en cada una de las posiciones se presentan un gran número de sustituciones, como mínimo tres sustituciones y como máximo 16. Así mismo observamos que en algunas posiciones se presentan más sustituciones en una rama en comparación con la otra. En la figura 20 se muestran las sustituciones en cada una de las ramas mediante diagramas de círculo, en los cuales observamos la frecuencia de ciertas sus sustituciones en cada una de las ramas (indicado por el tamaño de letra). Basándonos en esto encontramos que en algunas posiciones se observa la predominancia de una sustitución, mientras que en otras posiciones encontramos varias sustituciones que son más comunes; por otra parte también podemos observar que en algunas posiciones las sustituciones más comunes son por un residuo con la misma propiedad de polaridad, por ejemplo en ambas ramas, para la posición 56, la sustitución predominante es por un residuo no polar, y por lo contrario en las posiciones 116 y 167 las sustituciones son por residuos polares, mientras que en las posiciones 162 y 180 las sustituciones difieren en la polaridad entre la rama 1 y la rama 2; en la posición 162, las sustituciones en la rama 1 son por residuos polares, mientras que en la rama 2 las sustituciones más comunes son por residuos no polares; en las posición 180 las sustituciones en la rama 1 son por residuos no polar, y en la rama 2 es por residuos polares. En la posición 198 para la rama 1 las sustituciones más comunes incluyen residuos con distinta polaridad, polares, como la isoleucina y no polares como la

glutamina y serina; mientras que en la rama 2 la sustitución predominante es por un residuo polar.

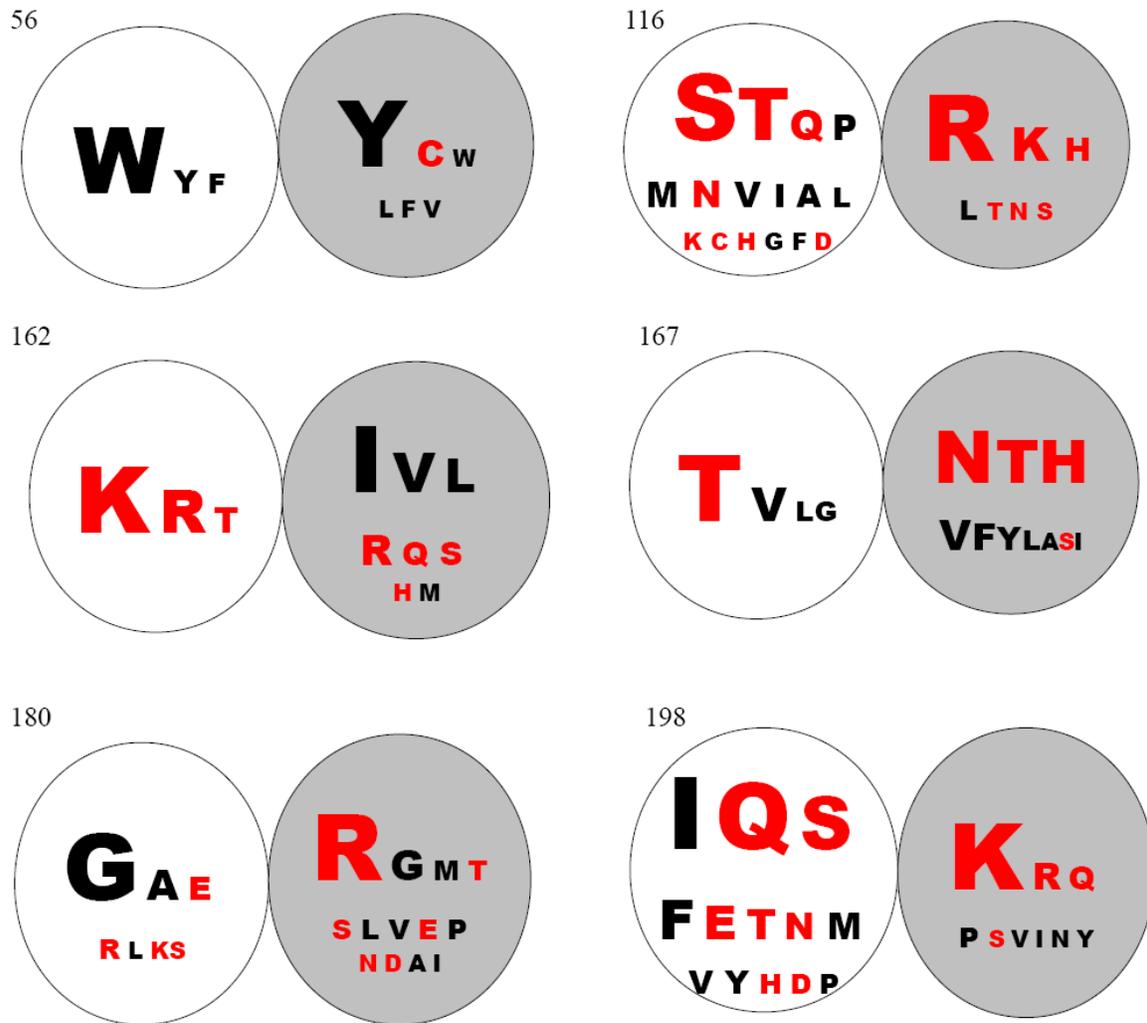


Figura 20. Representación mediante círculos de las sustituciones en las posiciones, 56, 116, 162, 167, 180 y 198, la numeración de acuerdo a la secuencias de eIF4E de *H. sapiens*. El círculo sin color, representa la rama 1, mientras que el círculo relleno en color gris representa la rama 2, el color de las letras indica las propiedades que comparten los residuos, el color rojo indica que son polares, mientras que el color negro indica que son no polares. El tamaño de la letra es proporcional al número de veces que se presenta ese residuo en cierta posición.

En las siguientes imágenes se muestran la distribución de las sustituciones a través del árbol enraizado (de la reconstrucción elaborada con el alineamiento de la región *core* que presentaba sustituciones en las posiciones 46, 69, 72, 73, 113 y 130) para cada una de las posiciones mostradas en la tabla 3.

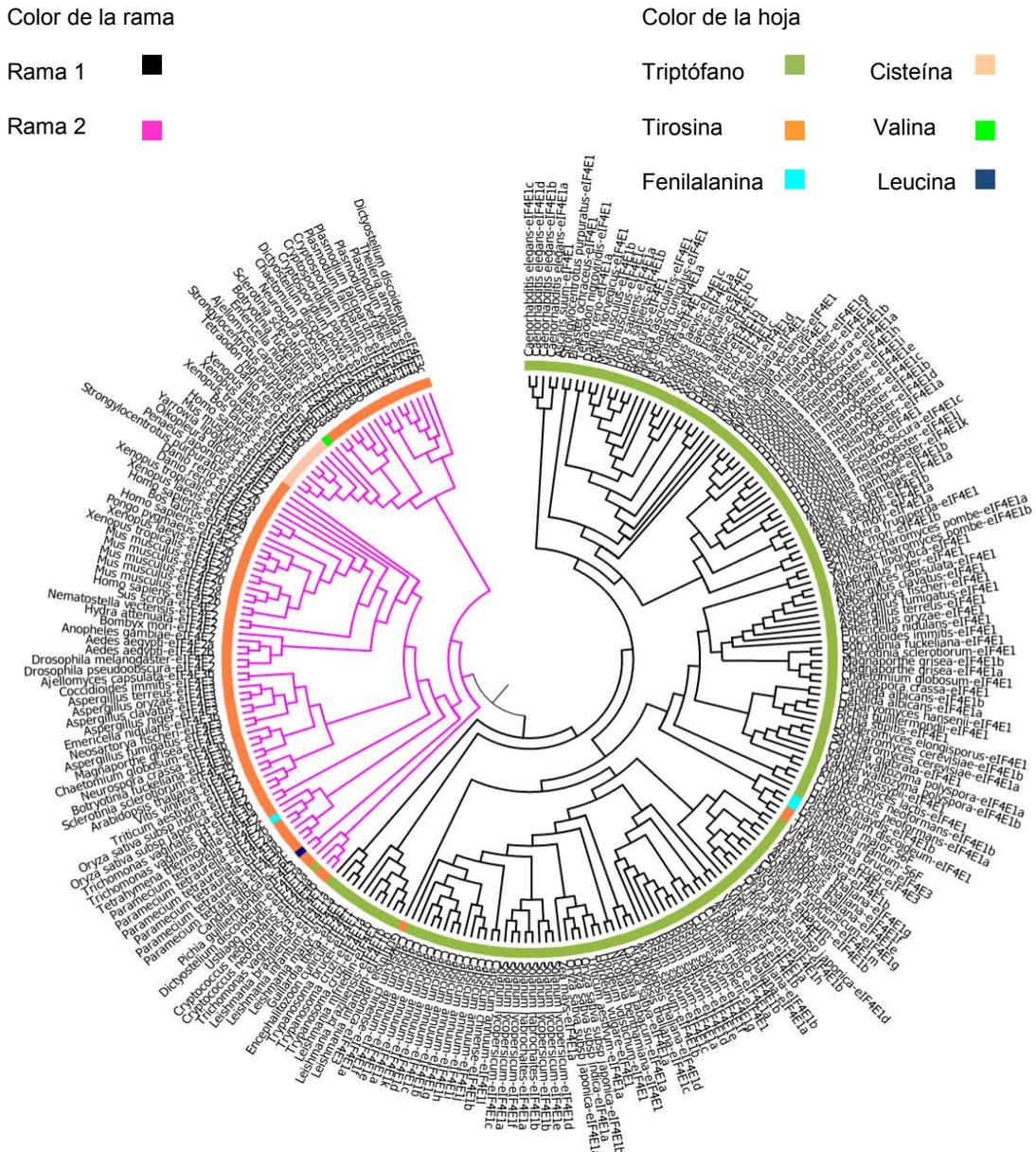


Figura 21. Distribución de las sustituciones que se presentan en la posición 56 de acuerdo a la numeración de eIF4E de *H. sapiens*, a través de la filogenia enraizada de la primera reconstrucción elaborada con el alineamiento de la región *core* que presentaba sustituciones en las posiciones 46, 69, 72, 73, 113 y 130; el color de la hoja indica que residuo se presentan en esa posición; mientras que el color de la rama indica a que rama pertenece, la “rama 1” incluye a todas las secuencias de la clase estructural I y 5 secuencias de la clase estructural III (4 de Euglenozoa y 1 de Rodophyta) mientras que la “rama 2” incluye a las secuencias de la clase estructural II y III.

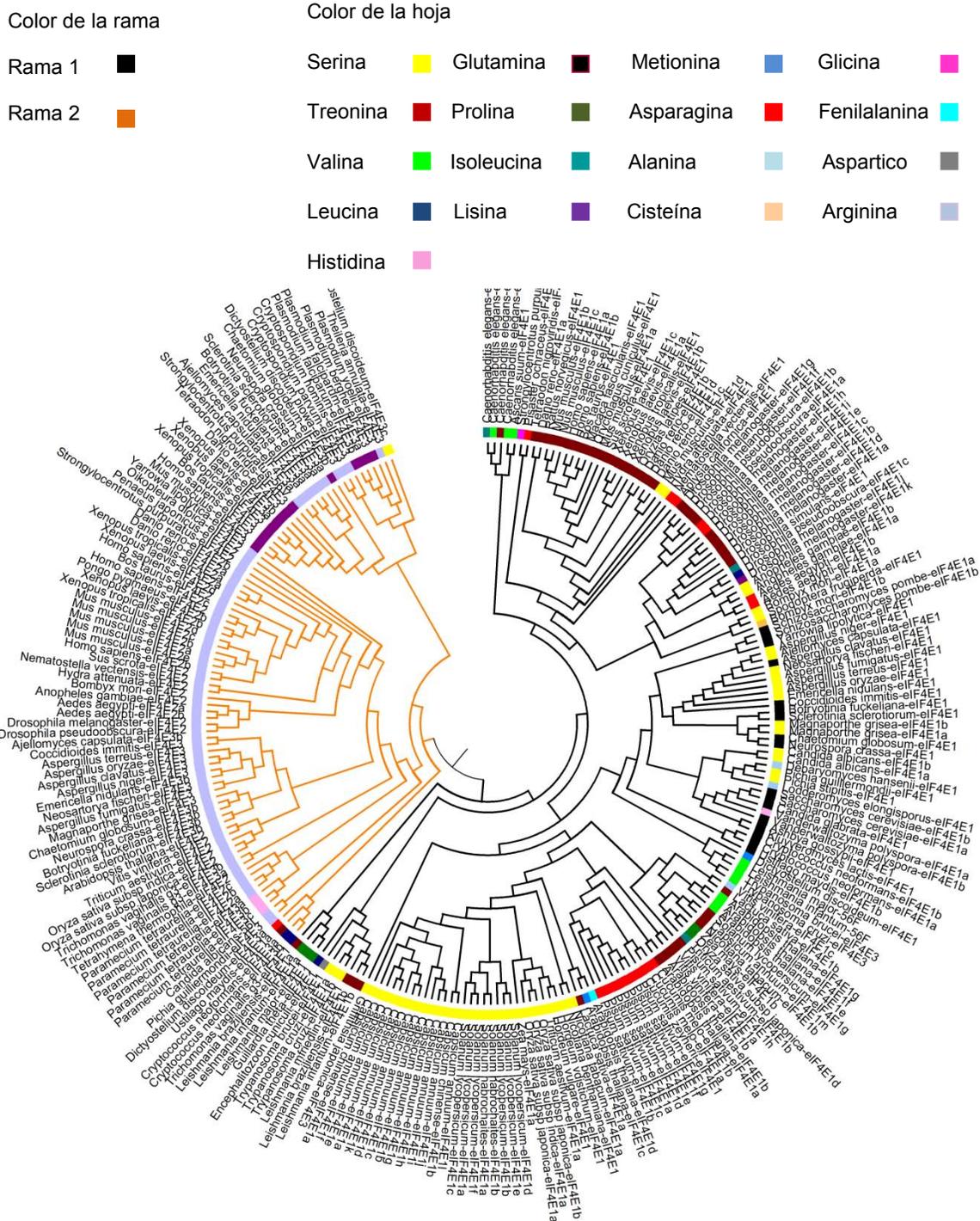


Figura 22. Distribución de las sustituciones que se presentan en la posición 116 de acuerdo a la numeración de eIF4E de *H. sapiens*, a través de la filogenia enraizada de la primera reconstrucción elaborada con el alineamiento de la región core que presentaba sustituciones en las posiciones 46, 69, 72, 73, 113 y 130; el color de la hoja indica que residuo se presentan en esa posición; mientras que el color de la rama indica a que rama pertenece, la “rama 1” incluye a todas las secuencias de la clase estructural I y 5 secuencias de la clase estructural III (4 de Euglenozoa y 1 de Rodophyta) mientras que la “rama 2” incluye a las secuencias de la clase estructural II y III.

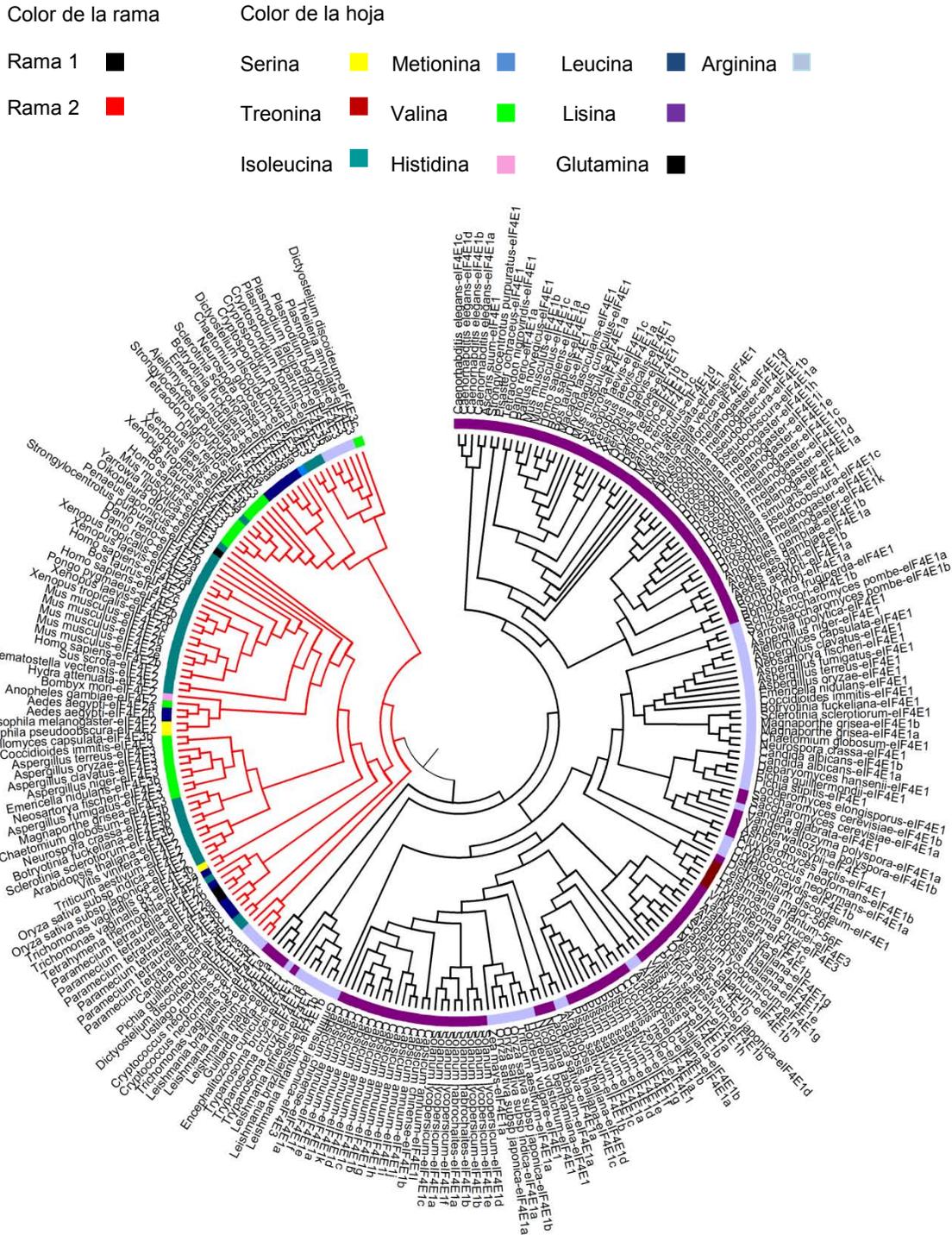


Figura 23. Distribución de las sustituciones que se presentan en la posición 162 de acuerdo a la numeración de eIF4E de *H. sapiens*, a través de la filogenia enraizada de la primera reconstrucción elaborada con el alineamiento de la región core que presentaba sustituciones en las posiciones 46, 69, 72, 73, 113 y 130; el color de la hoja indica que residuo se presenta en esa posición; mientras que el color de la rama indica a que rama pertenece, la “rama 1” incluye a todas las secuencias de la clase estructural I y 5 secuencias de la clase estructural III (4 de Euglenozoa y 1 de Rodophyta) mientras que la “rama 2” incluye a las secuencias de la clase estructural II y III.

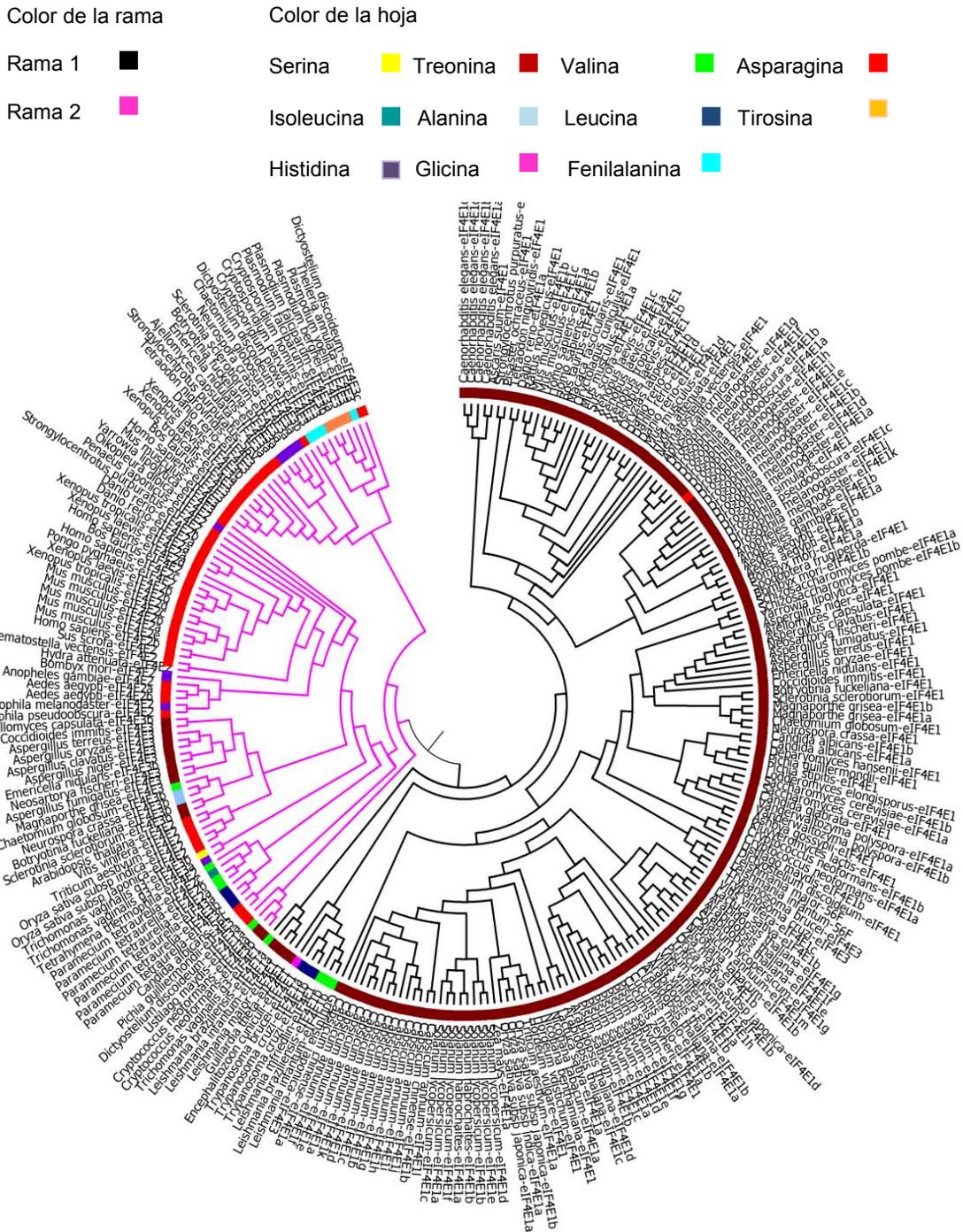


Figura 24. Distribución de las sustituciones que se presentan en la posición 167 de acuerdo a la numeración de eIF4E de *H. sapiens*, a través de la filogenia enraizada de la primera reconstrucción elaborada con el alineamiento de la región core que presentaba sustituciones en las posiciones 46, 69, 72, 73, 113 y 130; el color de la hoja indica que residuo se presentan en esa posición, aquellas secuencias sin forma de la hoja no presentan ningún residuo en esa posición; mientras que el color de la rama indica a que rama pertenece, la "rama 1" incluye a todas las secuencias de la clase estructural I y 5 secuencias de la clase estructural III (4 de Euglenozoa y 1 de Rodophyta) mientras que la "rama 2" incluye a las secuencias de la clase estructural II y III.

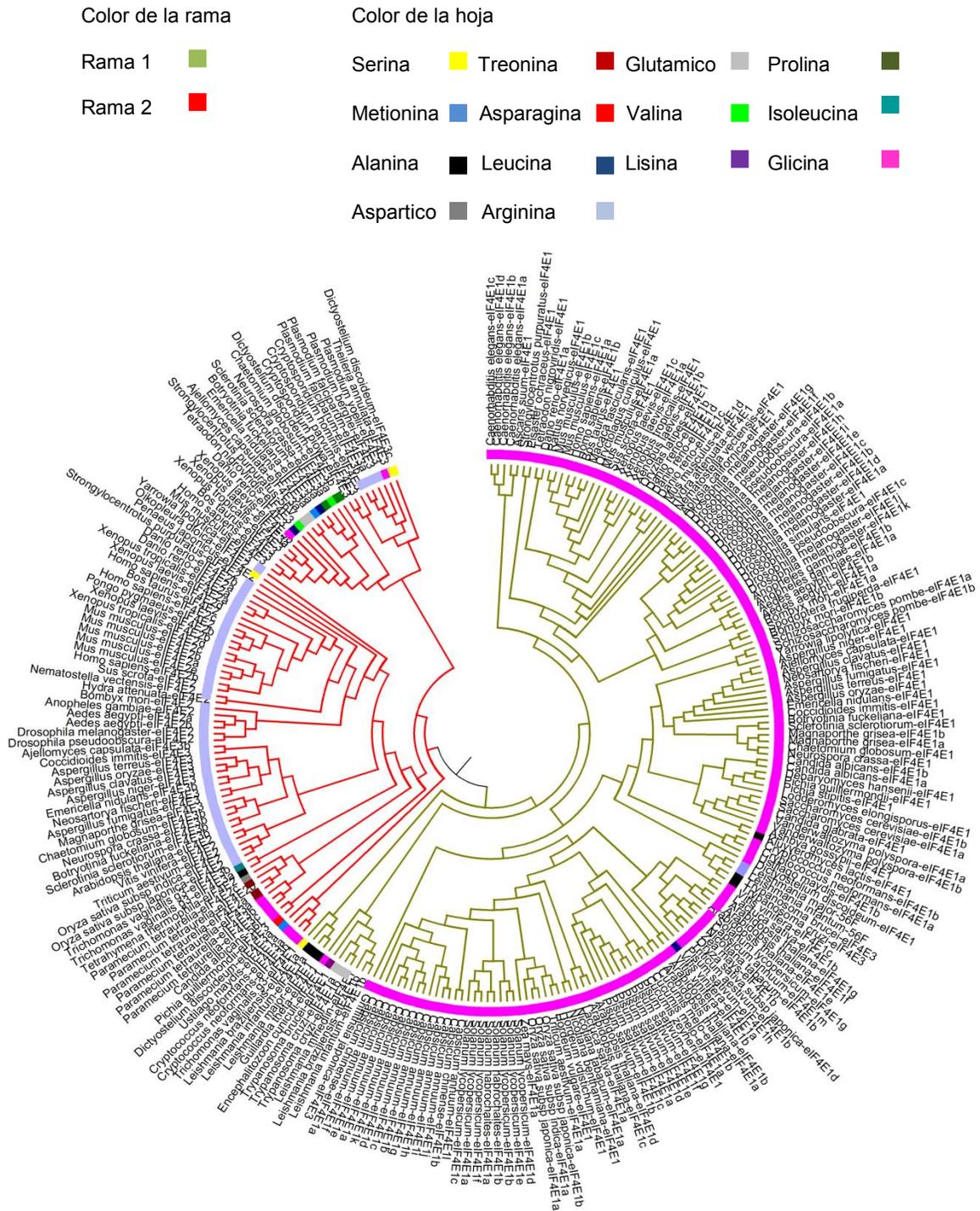


Figura 25. Distribución de las sustituciones que se presentan en la posición 180 de acuerdo a la numeración de eIF4E de *H. sapiens*, a través de la filogenia enraizada de la primera reconstrucción elaborada con el alineamiento de la región *core* que presentaba sustituciones en las posiciones 46, 69, 72, 73, 113 y 130; el color de la hoja indica que residuo se presentan en esa posición, aquellas secuencias sin forma de la hoja no presentan ningún residuo en esa posición; mientras que el color de la rama indica a que rama pertenece, la “rama 1” incluye a todas las secuencias de la clase estructural I y 5 secuencias de la clase estructural III (4 de Euglenozoa y 1 de Rodophyta) mientras que la “rama 2” incluye a las secuencias de la clase estructural II y III.

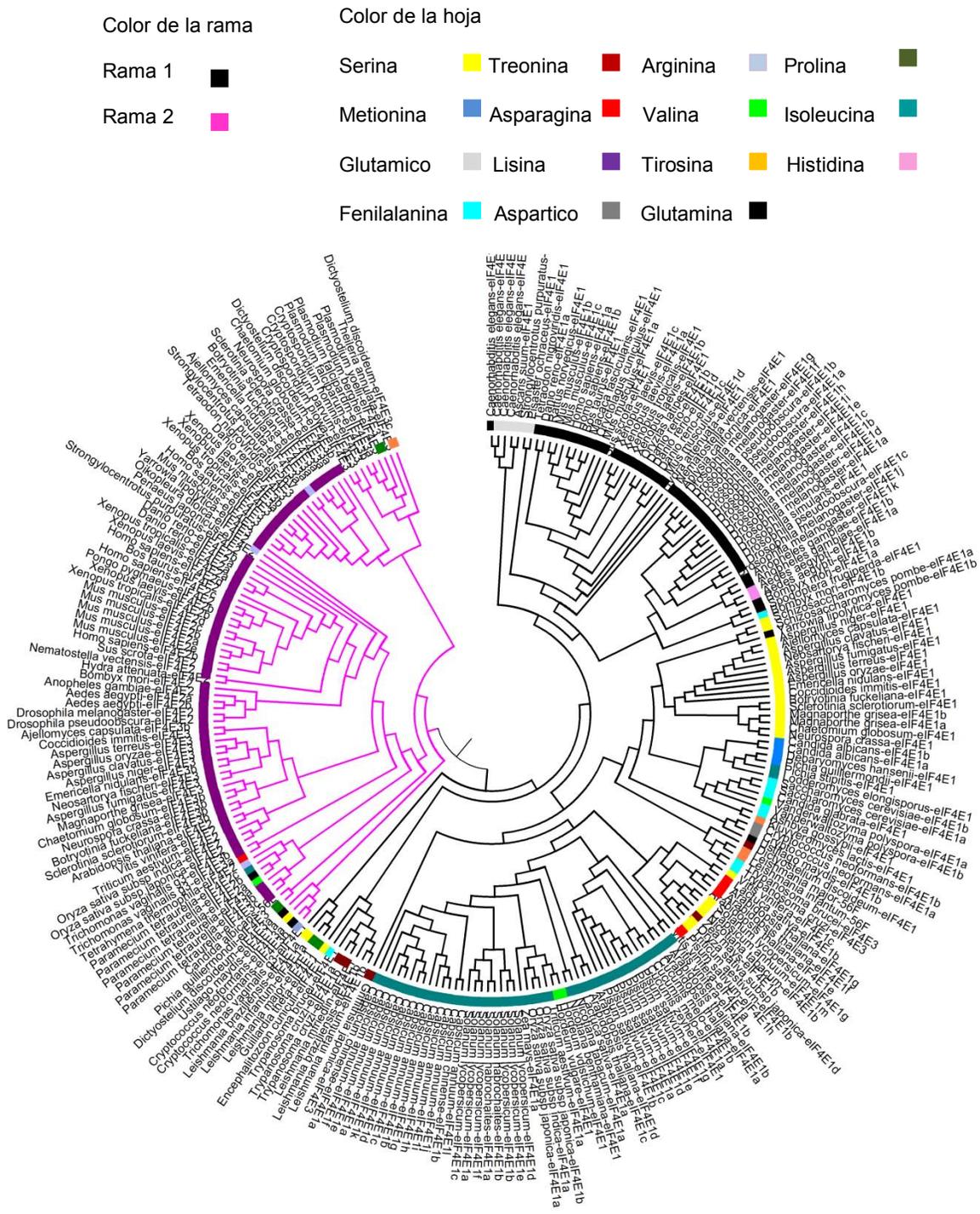


Figura 26. Distribución de las sustituciones que se presentan en la posición 198 de acuerdo a la numeración de eIF4E de *H. sapiens*, a través de la filogenia enraizada de la primera reconstrucción elaborada con el alineamiento de la región core que presentaba sustituciones en las posiciones 46, 69, 72, 73, 113 y 130; el color de la hoja indica que residuo se presentan en esa posición, aquellas secuencias sin forma de la hoja no presentan ningún residuo en esa posición; mientras que el color de la rama indica a que rama pertenece, la "rama 1" incluye a todas las secuencias de la clase estructural I y 5 secuencias de la clase estructural III (4 de Euglenozoa y 1 de Rodophyta) mientras que la "rama 2" incluye a las secuencias de la clase estructural II y III.

Observando las figuras 21-26, podemos notar que la principal característica que permanece constante a través de las diversas posiciones es la alta variación en sustituciones que se presentan en una de las ramas, mientras que en la otra se observan una menor variación ya presenta uno o dos residuos como las sustituciones con mayor frecuencia. Una posible explicación es que esto se deba a que las ramas se originaron de un evento de duplicación, liberando a uno de los duplicados resultantes de ciertas limitaciones estructurales o funcionales permitió sustituciones en ciertas posiciones.

Por otra parte, también observamos que la rama 2, que incluye a secuencias pertenecientes a la clase estructural II y III, se puede subdividir en 2 ramas; la rama 2a, contiene a las secuencias pertenecientes a la clase estructural II de Metazoa, Viridiplantae y Alveolata, y secuencias de la clase III de Fungi y Alveolata. La rama 2b, solamente presenta secuencias de la clase estructural III de los Metazoa, los Alveolata y los Fungi. Al igual que las ramas principales, se observa sustituciones características en ciertas posiciones, y que se muestran en la tabla 4, la numeración es de acuerdo a la secuencia de eIF4E de *H. sapiens*

Tabla 4. Sustituciones en las posiciones que pueden ser empleadas para diferenciar las sub-ramas secundarias de la rama 2, las posiciones son de acuerdo a la numeración de la secuencia de eIF4E de *H. sapiens*.

Posición en la secuencia	Rama 2a	Rama 2b
43	W Y F L	W
70	E Q R D	Q E G A
81	R K H L	Q P L S D N
114	I V T L Q	K T E H R V D
117	L F V C	V I A F C
131	E A L	N K T H L
153	V I S A	S G R V

En la tabla 4 se observa que hay diversas sustituciones que pueden presentarse en cada una de las posiciones. En la figura 27 se muestran las sustituciones en cada una de las sub-ramas mediante diagramas de círculo, en los cuales observamos la frecuencia de ciertas sustituciones indicado por el tamaño de letra; también podemos observar que en ambas ramas, en la mayoría de las posiciones la sustituciones más comunes mantienen la misma propiedad. La excepción a esto son las posiciones 114 y 153, en las cuales la sustitución más común en la sub-rama 2a es por un residuo no polar, mientras que en la rama 2b el residuo es polar. Por otra parte, en la posición 81 notamos que en la sub-rama 2a el residuo predominante es polar, y en la sub-rama 2b las sustituciones más comunes son por residuos con propiedades diferentes, polar y no polar, lo cual podría indicar que en esta sub-rama la propiedad del residuo no es importante, y por lo tanto no se conserva.

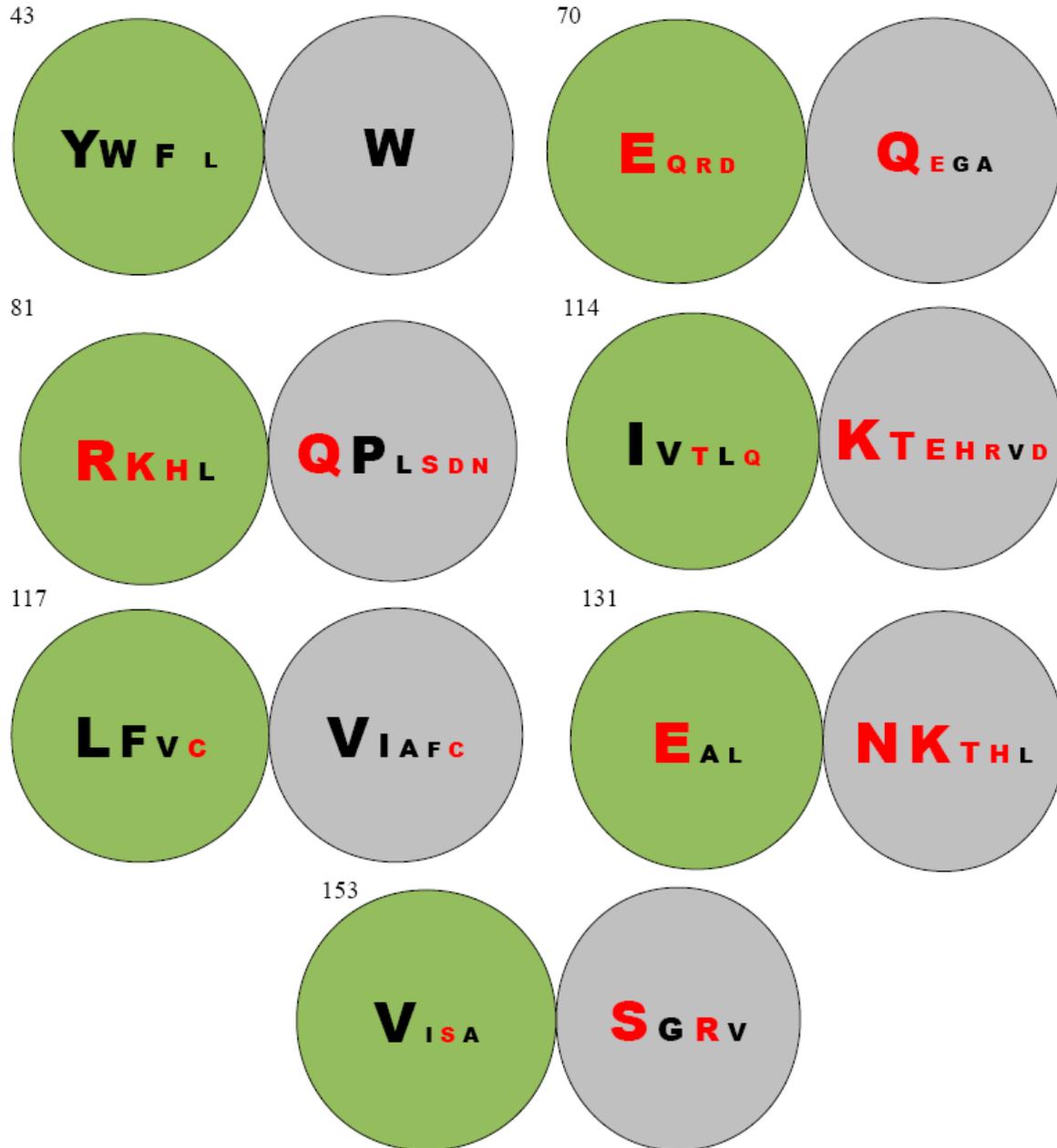


Figura 27. Representación mediante círculos de las sustituciones en las posiciones, 43, 70, 81, 114, 117, 131 y 153, la numeración es de acuerdo a la secuencias de eIF4E de *H. sapiens*. El color del círculo indica la sub-rama, el color verde representa la sub-rama 2a (secuencias pertenecientes a la clase estructural II de Metazoa, Viridiplantae y Alveolados, y secuencias de la clase III de Fungi y Alveolados), el color gris representa a la sub-rama 2b (secuencias de la clase estructural III de Metazoa, Alveolados y Fungi); el color de las letras indica las propiedades que comparten los residuos, el color rojo indica que son polares, mientras que el color negro indica que son no polares. El tamaño de la letra es proporcional al número de veces que se presenta ese residuo en cierta posición.

En la siguiente imagen se muestran la distribución de las sustituciones en las posiciones mencionadas en la tabla 4 a través de la rama 2, tomada del árbol enraizado de la primera reconstrucción filogenética, construido con el alineamiento de secuencias de la región core que presenta sustituciones en las posiciones 46, 69, 72, 73, 113 y 130

A) Posición 43

Color de la rama

Rama 2a ■

Rama 2b ■

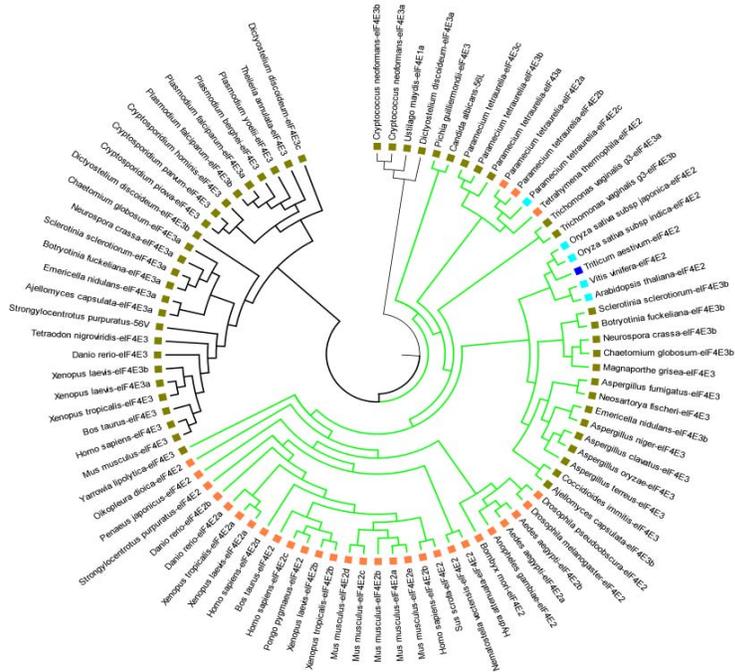
Color de la hoja

Triptófano ■

Tirosina ■

Leucina ■

Fenilalanina ■



B) Posición 70

Color de la rama

Rama 2a ■

Rama 2b ■

Color de la hoja

Glutamico ■

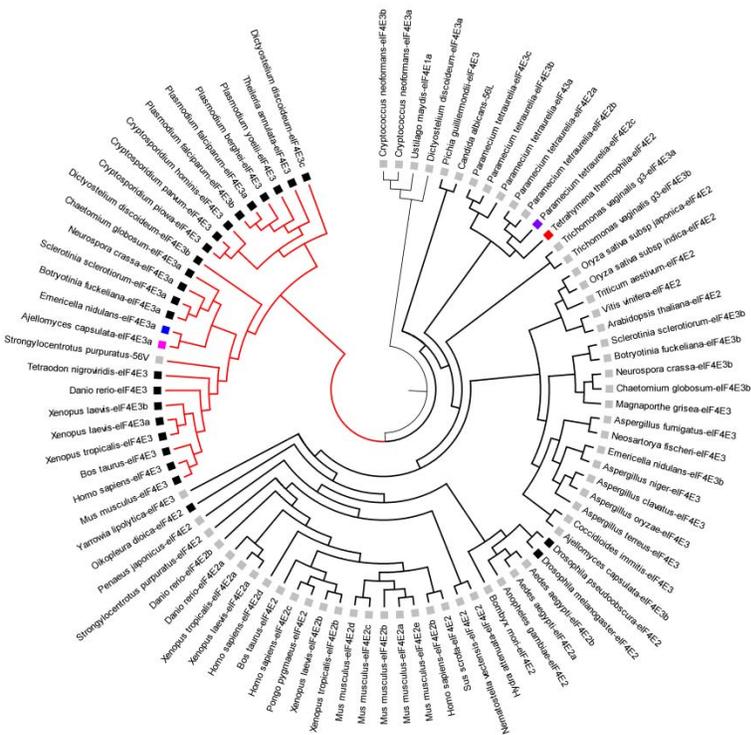
Aspartico ■

Glutamina ■

Arginina ■

Glicina ■

Alanina ■



C) Posición 81

Color de la rama

Rama 2a

Rama2b

Color de la hoja

Arginina

Lisina

Histidina

Leucina

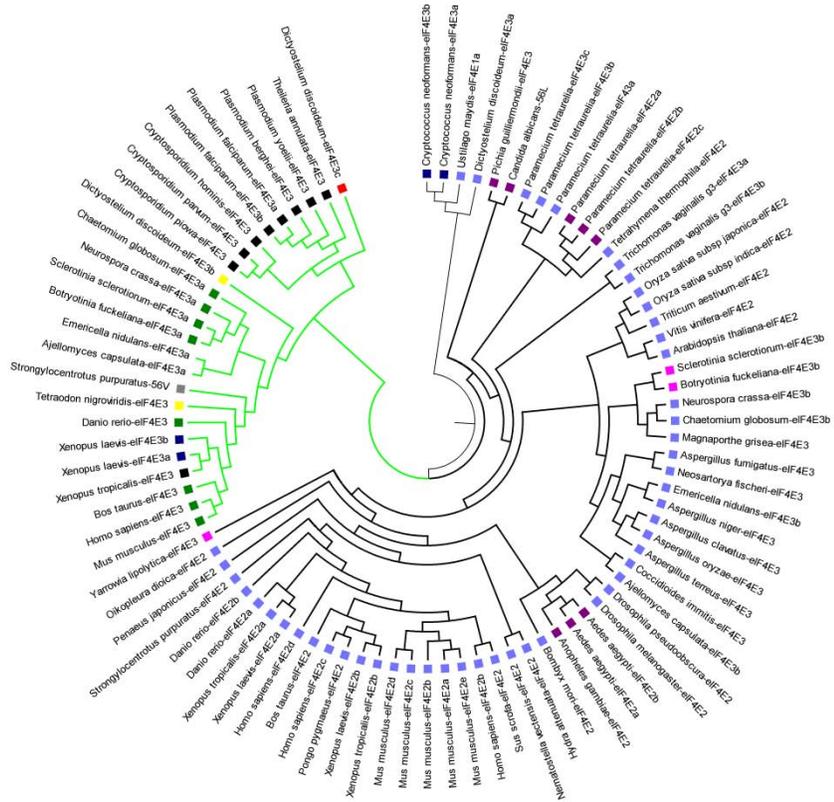
Prolina

Glutamina

Serina

Aspartico

Asparagina



D) Posición 114

Color de la hoja

Isoleucina

Valina

Treonina

Leucina

Glutamina

Lisina

Glutamico

Histidina

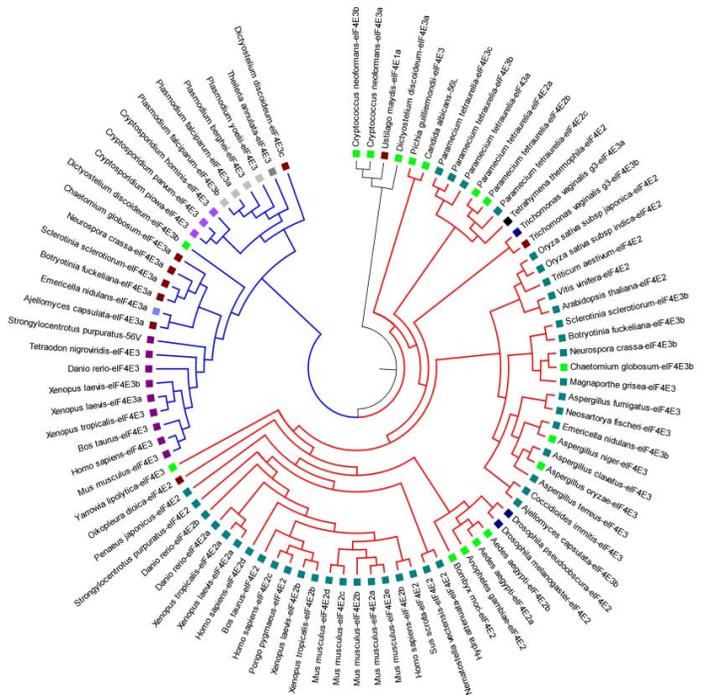
Arginina

Aspartico

Color de la rama

Rama 2a

Rama2b



E) Posición 117

Color de la rama

Rama 2a ■

Rama2b ■

Color de la hoja

Leucina ■

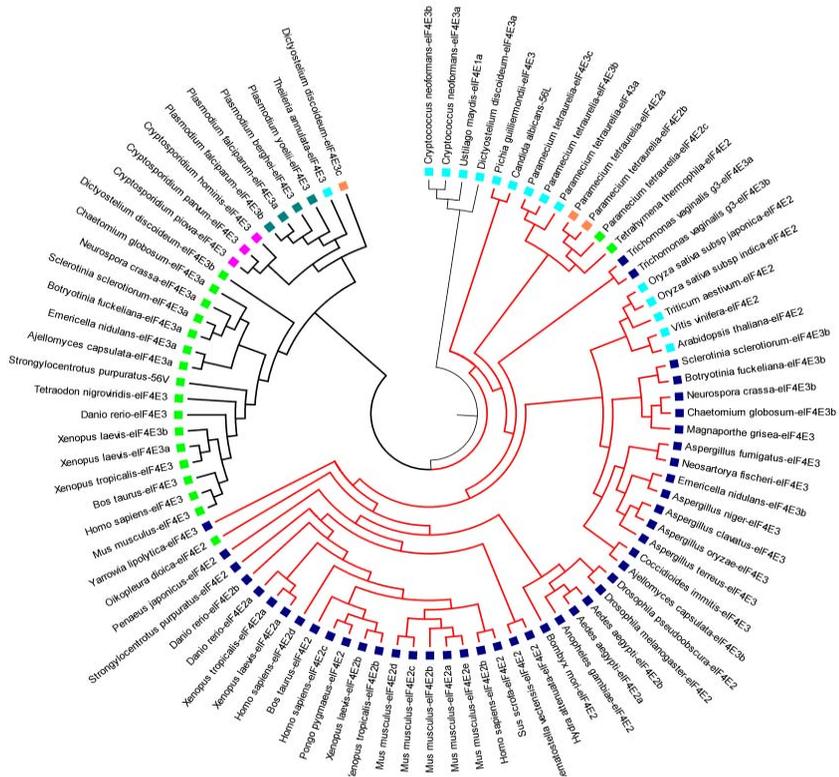
Fenilalanina ■

Valina ■

Cisteína ■

Isoleucina ■

Alanina ■



F) Posición 131

Color de la rama

Rama 2a ■

Rama2b ■

Color de la hoja

Glutamico ■

Alanina ■

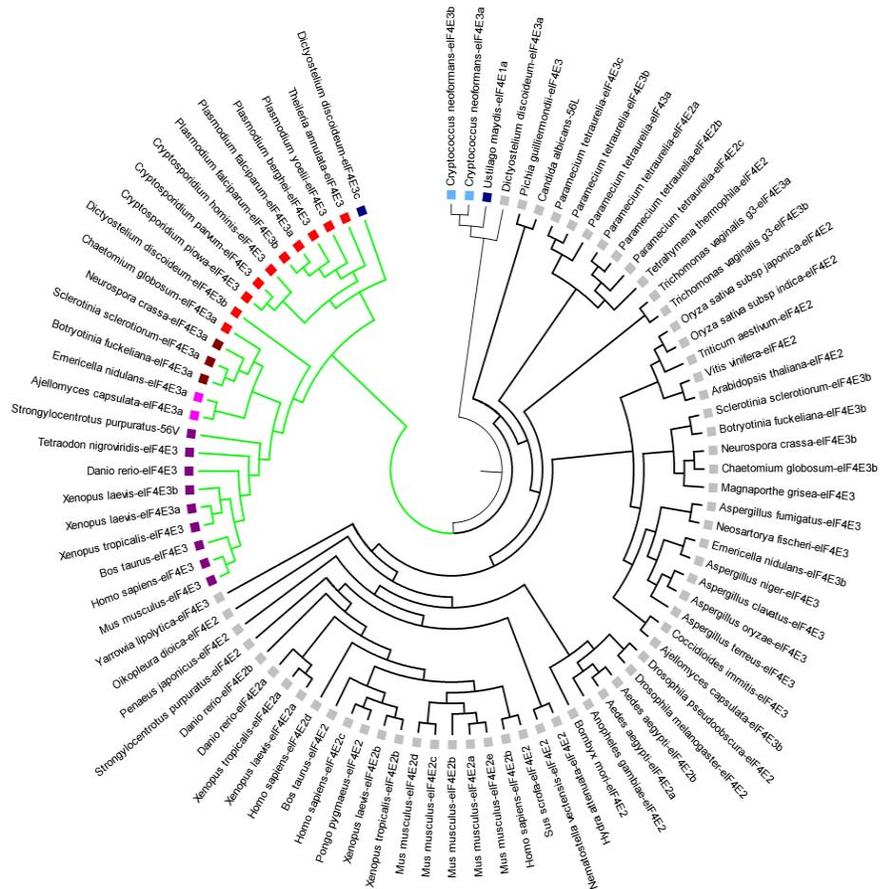
Leucina ■

Asparagina ■

Lisina ■

Treonina ■

Histidina ■



G) Posición 153

Color de la rama

Rama 2a

Rama2b

Color de la hoja

Valina

Isoleucina

Serina

Alanina

Glicina

Arginina

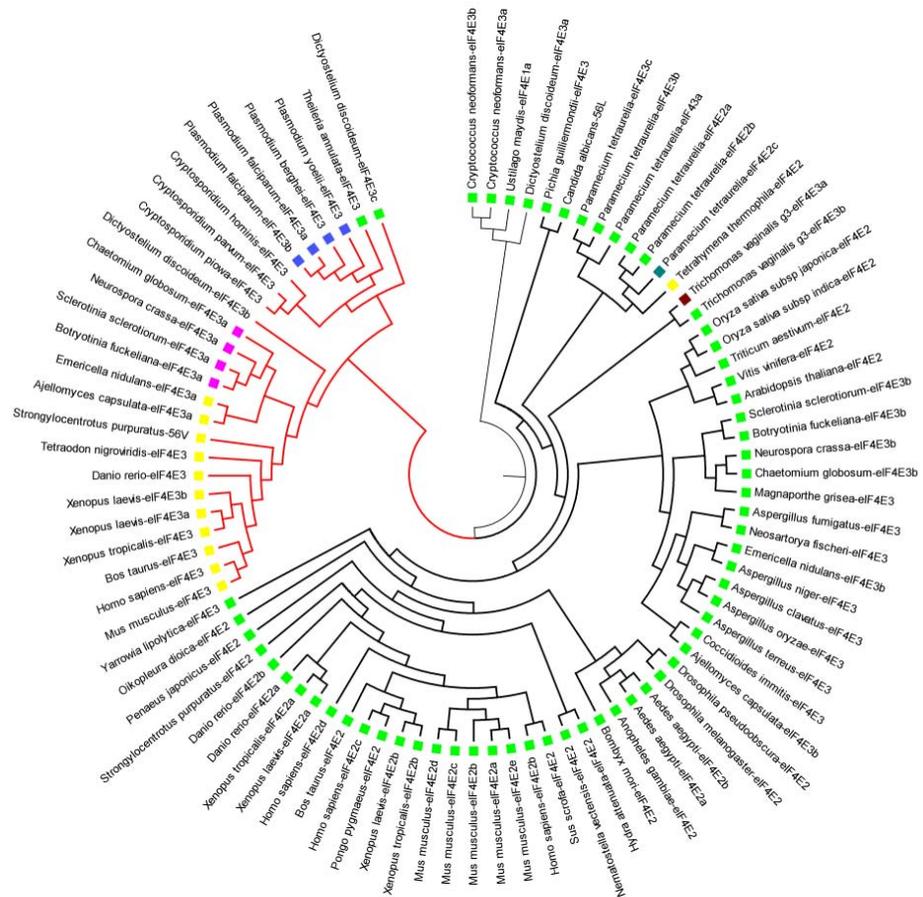
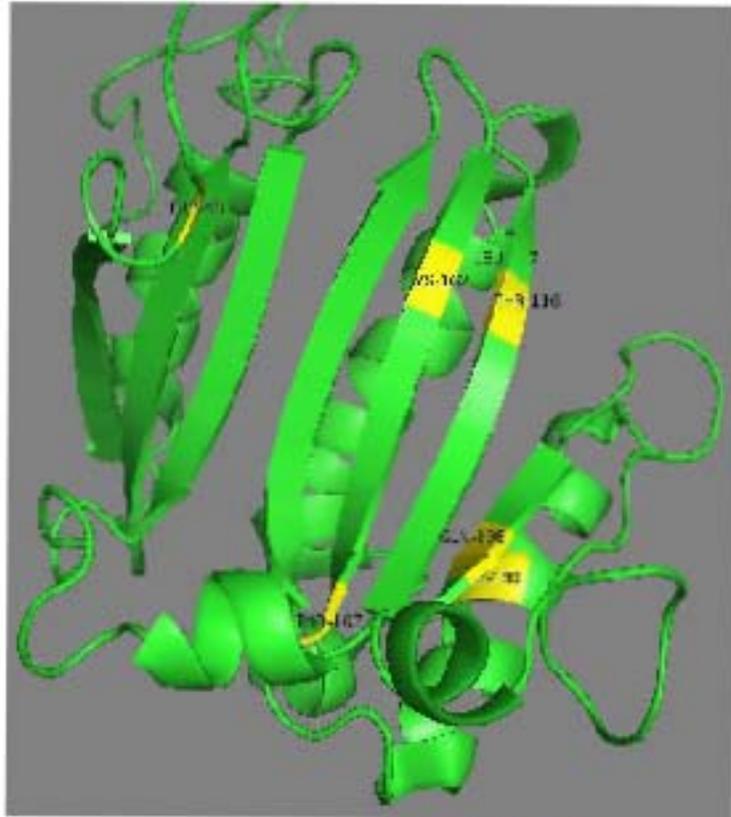


Figura 28. Distribución de las sustituciones a través de la rama 2 tomada del árbol enraizado de la primera reconstrucción filogenética, que se construyó con el alineamiento de secuencias de la región *core* que presentaba sustituciones en las posiciones 46, 69, 72, 73, 113 y 130; el color de la rama indica a que rama secundaria pertenece Rama 2a (secuencias pertenecientes a la clase estructural II de Metazoa, Viridiplantae y Alveolados, y secuencias de la clase III de Fungi y Alveolados), Rama 2b (secuencias de la clase estructural III de Metazoa, Alveolados y Fungi), el color de la hoja indica los residuos que se presentan en la posición, aquellas ramas que no presentan hojas, es porque no presenta ningún residuo en esa posición del alineamiento. Las posiciones se indican en A) posición 43, B) posición 70, C) posición 81, D) posición 114, E) posición 117, F) posición 131 y G) posición 153, el número de las posiciones es de acuerdo a la numeración de eIF4E de *H. sapiens*.

Al igual que en las ramas principales (Rama1 y Rama2), en las ramas secundarias (Rama 2a y Rama 2b) observamos una alta variación de sustituciones, mientras que en la otra rama la variación es menor y predominan una o dos sustituciones. Posiblemente esta característica sea resultado de un evento de duplicación.

En las figuras 29 y 30 se muestran las posiciones para la diferenciación de las ramas principales y para las ramas secundarias, en un diagrama de listones 3D para eIF4E de *H. sapiens*.

A)



B)

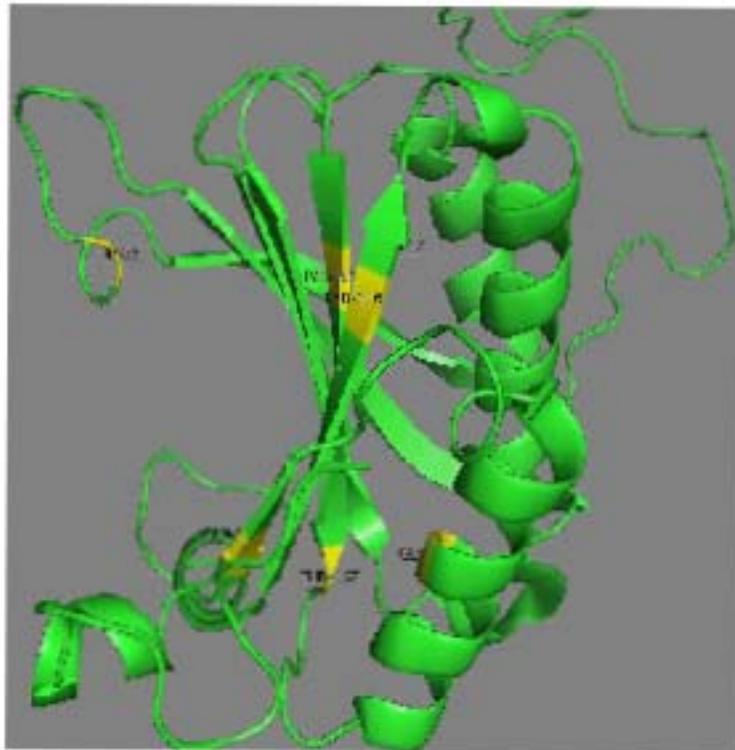


Figura 29. Distribución de las posiciones presentadas en la tabla 3 en un diagrama de listones 3D de eIF4E de *H. sapiens*. En amarillo se muestran las posiciones 56, 116, 162, 167, 180, 198, empleadas para diferenciar las ramas principales. **A)** Vista frontal de la parte cóncava de eIF4E, **B)** Vista lateral de eIF4E.

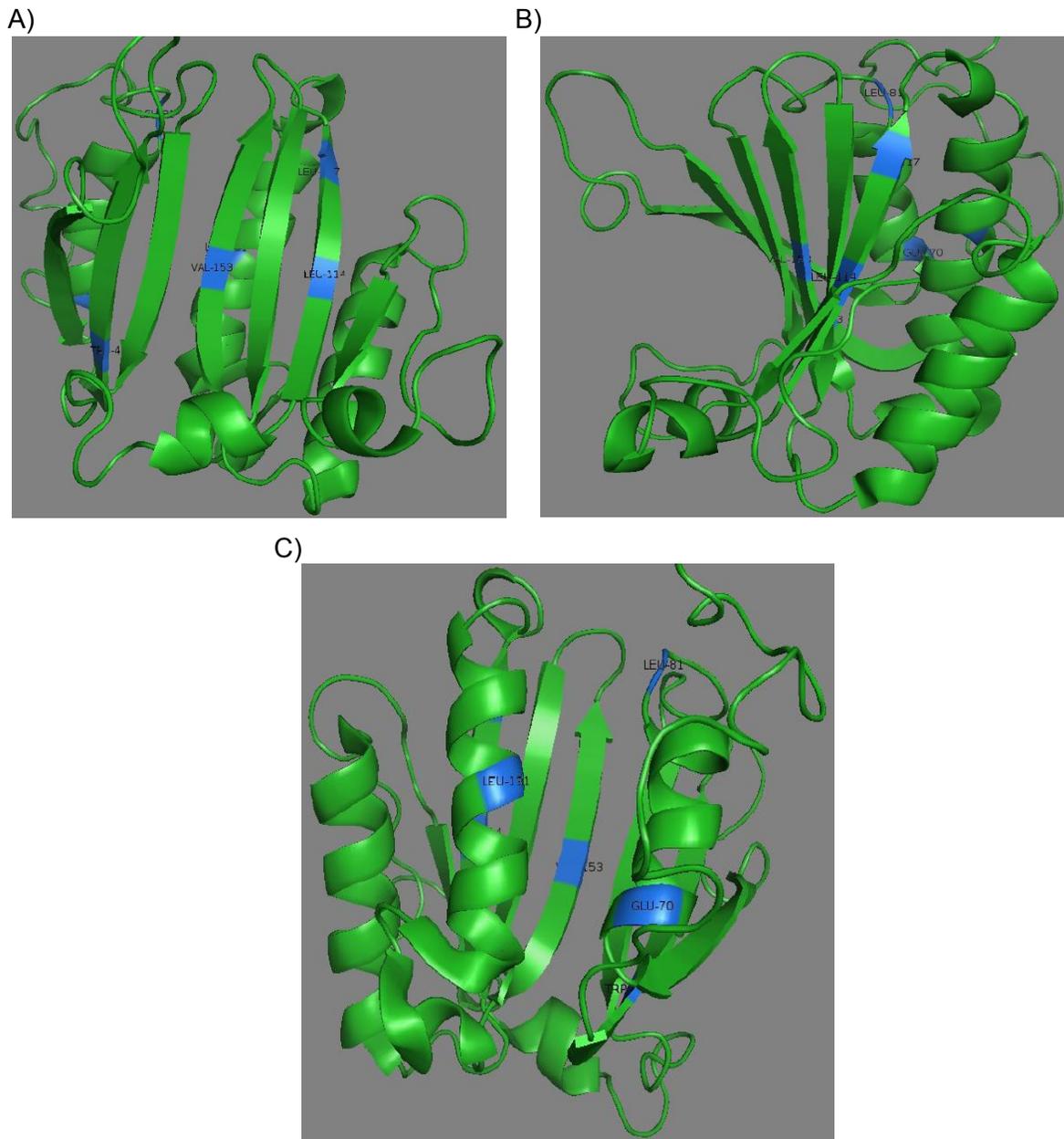


Figura 30. Distribución de las posiciones presentadas en la tabla 4 a través de un diagrama de listones 3D de eIF4E de *H. sapiens*. En azul se muestran las posiciones 43, 70, 81, 114, 17, 131, 153, empleadas para diferenciar las ramas secundarias. **A)** Vista frontal de la parte cóncava de eIF4E, **B)** Vista lateral de eIF4E y **C)** Vista frontal de la parte convexa de eIF4E.

De acuerdo a las figuras 29 y 30 notamos que la distribución de las posiciones que se muestran en la tabla 3, ocurren mayoritariamente en la parte cóncava de la proteína, mientras que las posiciones de la tabla 4 se ubican tanto en la parte cóncava como en la convexa de eIF4E. No sabemos si aquellos residuos ubicados en la parte cóncava de la proteína tengan alguna interacción con el *cap*, pero por su ubicación cabe la posibilidad

que esta interacción pueda llevarse a cabo. Con respecto a las posiciones ubicadas en la parte convexa, la posición 70 es la única de ellas que participa en la interacción con 4G y las 4EBP's.

5.3 Detección de Sitios bajo Selección Natural.

La detección de selección natural, se realizó con el programa GA-Branch (Kosakovsky-Pond et al 2005) el cual detecta selección a nivel de ramas utilizando el árbol enraizado de la primera reconstrucción filogenética (figura 9), ya que incluye un mayor número de secuencias y de diversos linajes. El modelo con el cual se detectó la presencia de selección utilizó 5 clases de tasas dN/dS (tasas de sustitución no sinónimas/ tasas de sustitución sinónimas) indicando que las ramas pueden ser clasificadas dentro de estas 5 clases de acuerdo a la dN/dS que presenten.

De las 5 clases de dN/dS que el modelo empleó, la que nos interesa es la clase que presenta una tasa de dN/dS de 8.1451, ya que indica que las ramas agrupadas dentro de esta clase presentan selección natural positiva. Con respecto a las 4 clases restantes de dN/dS que utilizó el modelo, claramente nos indican que estas ramas están bajo selección natural negativa. Sin embargo, la presión que ejerce la selección sobre cada una de las clases puede variar, como se observa en la clase que presenta una dN/dS de 0.7541 que es ligeramente menor a 1. Por lo tanto, las ramas agrupadas dentro de esta clase presentan sustituciones no sinónimas, sin embargo son menores en comparación con las sustituciones sinónimas; por otra parte observamos que en las demás clases su dN/dS (0.0953, 0.0223 y 0.0014) es muy cercana a 0 por lo tanto la selección negativa está actuando con una mayor presión, ya que las sustituciones no sinónimas son mucho menos frecuentes y predominan más las sustituciones sinónimas. En la figura 31 se muestra la distribución de las ramas que se agrupan dentro de las 5 clases de dN/dS a través del árbol enraizado representado en forma circular de la primera reconstrucción filogenética.

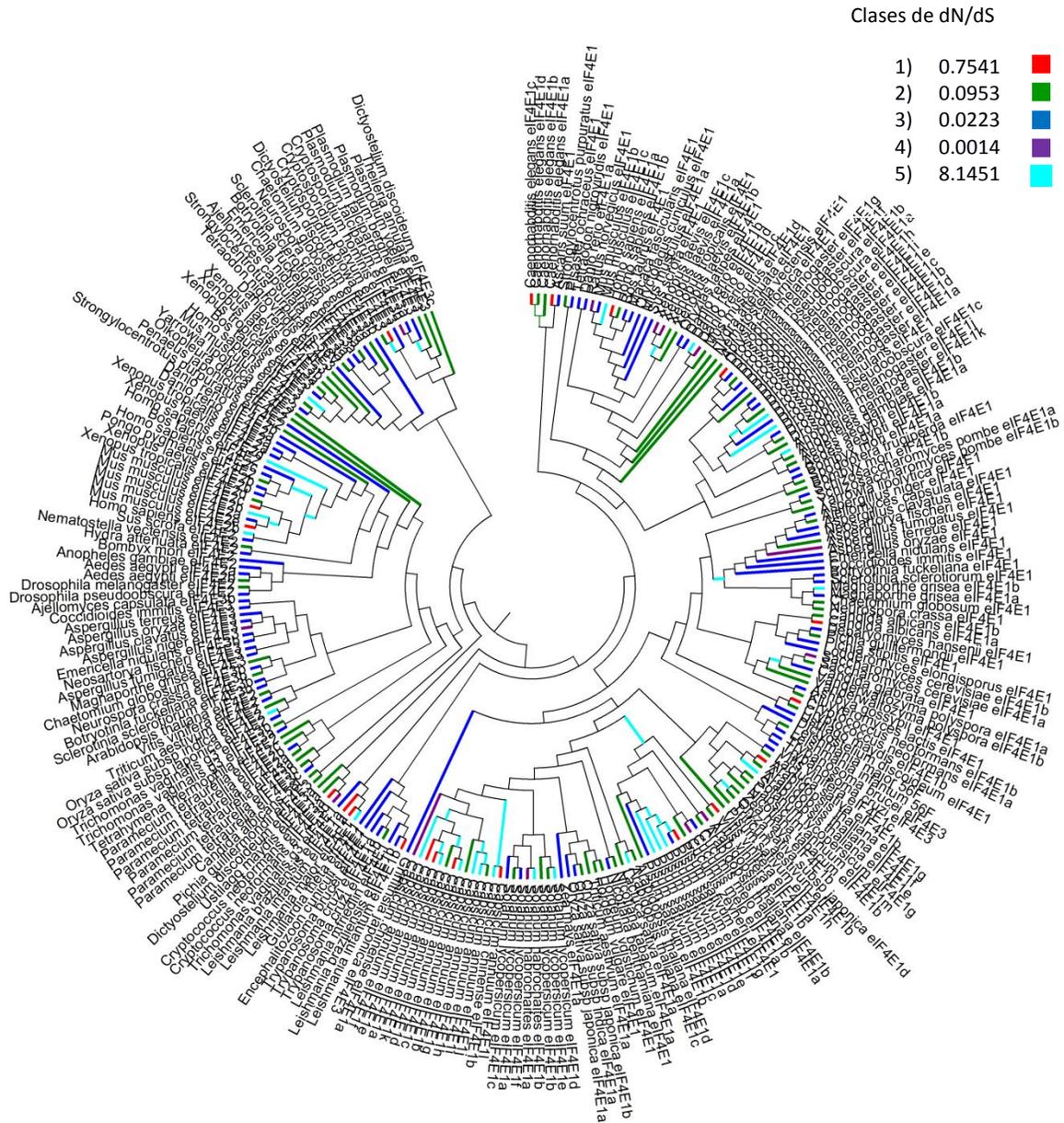


Figura 31. Distribución de las ramas agrupadas dentro de las 5 clases de dN/dS en la primera reconstrucción filogenética enraizada representada en forma circular. El color de la rama indica a que clase pertenecen, las ramas internas solo son coloreadas cuando pertenecen a la clase que presenta una dN/dS de 8.1451 ya que están bajo selección natural positiva.

En la figura anterior podemos observar como las ramas pertenecientes a las 5 clases de dN/dS se distribuyen a través de todo el árbol, incluyendo miembros de las 3 clases estructurales. Sin embargo, existe un mayor número de ramas de la clase dN/dS con valor de 8.1451 en miembros de la clase estructural I. Algunos ejemplos de ello se muestran en la figura 32.

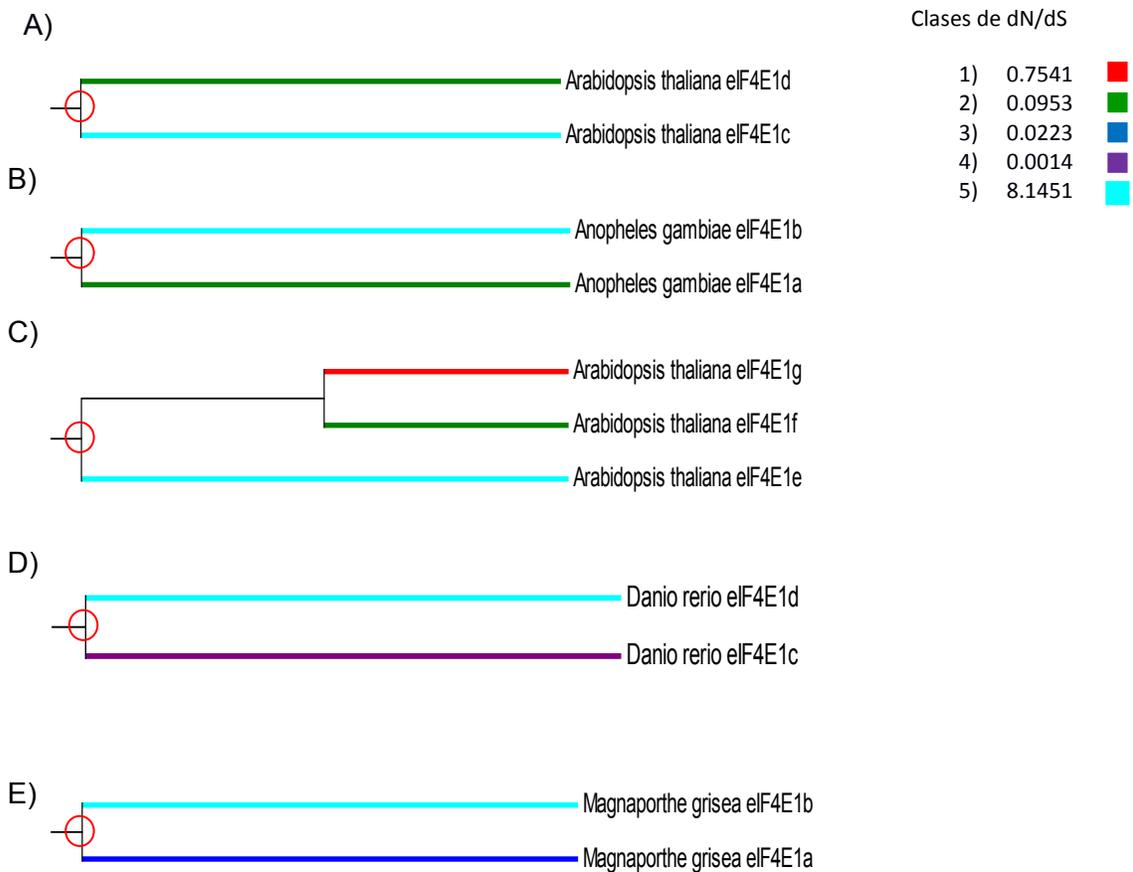


Figura 32. Se muestran diversos subárboles tomados de la primera filogenia enraizada, el color de la rama indica a qué clase de dN/dS pertenece. En A, B, C D y E después de un evento de duplicación sólo hay una rama con selección positiva interna o terminal. El círculo rojo indica un evento de duplicación del cual uno de los duplicados resultantes presenta selección natural positiva

En la figura anterior (32) se muestran diversos subárboles, en los cuales comparten la característica de pertenecer a la clase estructural I y presentar un evento de duplicación, y posterior a este evento en uno de los duplicados resultantes se observa evidencia de eventos de selección natural positiva detectada mediante pruebas de rama. Por ello estas ramas se agrupan en la clase que presenta una dN/dS de 8.1451, que es lo que se espera después de un evento de duplicación. Uno de los duplicados resultantes se libera de las restricciones funcionales permitiendo que el aminoácido cambie en algunas posiciones importantes para la función, con lo cual la tasa de sustitución no sinónima sea mayor que la tasa de sustitución sinónima. Observamos en los subárbol de la figura 32, que después del evento de duplicación uno de los duplicados resultantes presenta un valor de dN/dS > 1, mientras que el otro una dN/dS < 1.

Un caso distinto al anterior podría observarse en la figura 33 en donde se muestra tres subárboles en los cuales secuencias de la misma clase estructural y que pertenecen al mismo organismo que se originaron de varios eventos de duplicación, solo una rama terminal presenta selección natural positiva, con lo cual podemos inferir que los duplicados que presentan selección negativa son aun importantes funcionalmente.

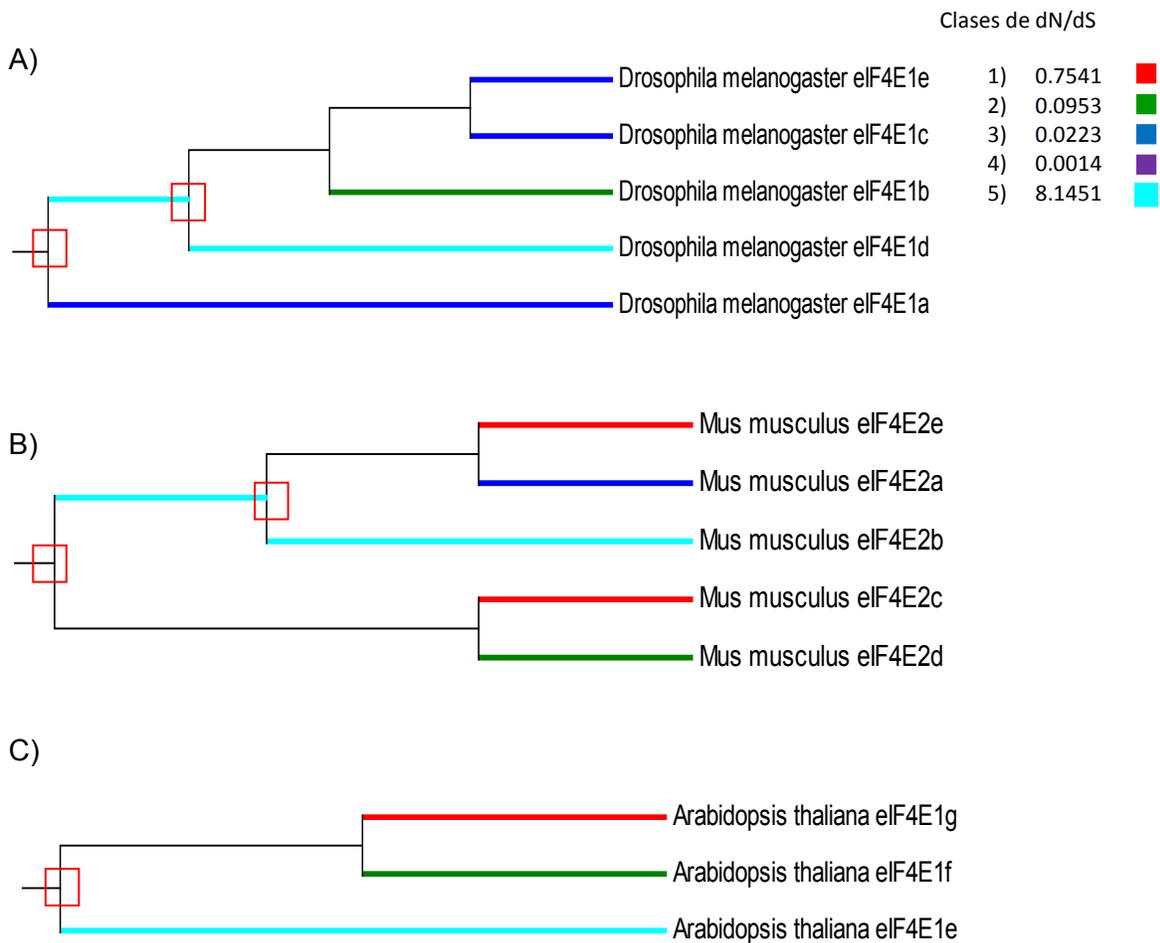


Figura 33. Se muestran dos subárboles tomados de la primera filogenia enraizada, el color de la rama indica a qué clase de dN/dS pertenece. En A) subárbol que agrupa las secuencias de eIF4E de la clase estructural I de *Drosophila melanogaster*; en B) subárbol que agrupa las secuencias de de la clase estructural II de *Mus musculus* y en C) subárbol que agrupa secuencias de la clase estructural I de *Arabidopsis thaliana*. El recuadro rojo indica un evento de duplicación del cual uno de los duplicados resultantes presenta selección natural positiva.

Por otra parte puede haber casos contrarios al presentado en la figura 33. Secuencias que se originaron de varios eventos de duplicación que pertenecen al mismo organismo y

a la misma clase estructural presentan selección natural positiva, es lo que se espera. Sin embargo, lo sobresaliente de estos casos es que la mayoría pertenecen a miembros del linaje orgánico de Viridiplantae y especialmente en aquellos que son atacados por virus (potyvirus). Estudios han demostrado que la resistencia contra estos virus se debe a la sustitución de unos cuantos aminoácidos en eIF4E. Estas sustituciones se localizan en dos regiones cercanas al sitio de unión a *cap* en la superficie de la proteína (Robaglia y Caranta, 2006). En la figura 34 se muestran los subárboles que agrupan las secuencias de la clase estructural I de *P. sativum* y *C. annuum*.

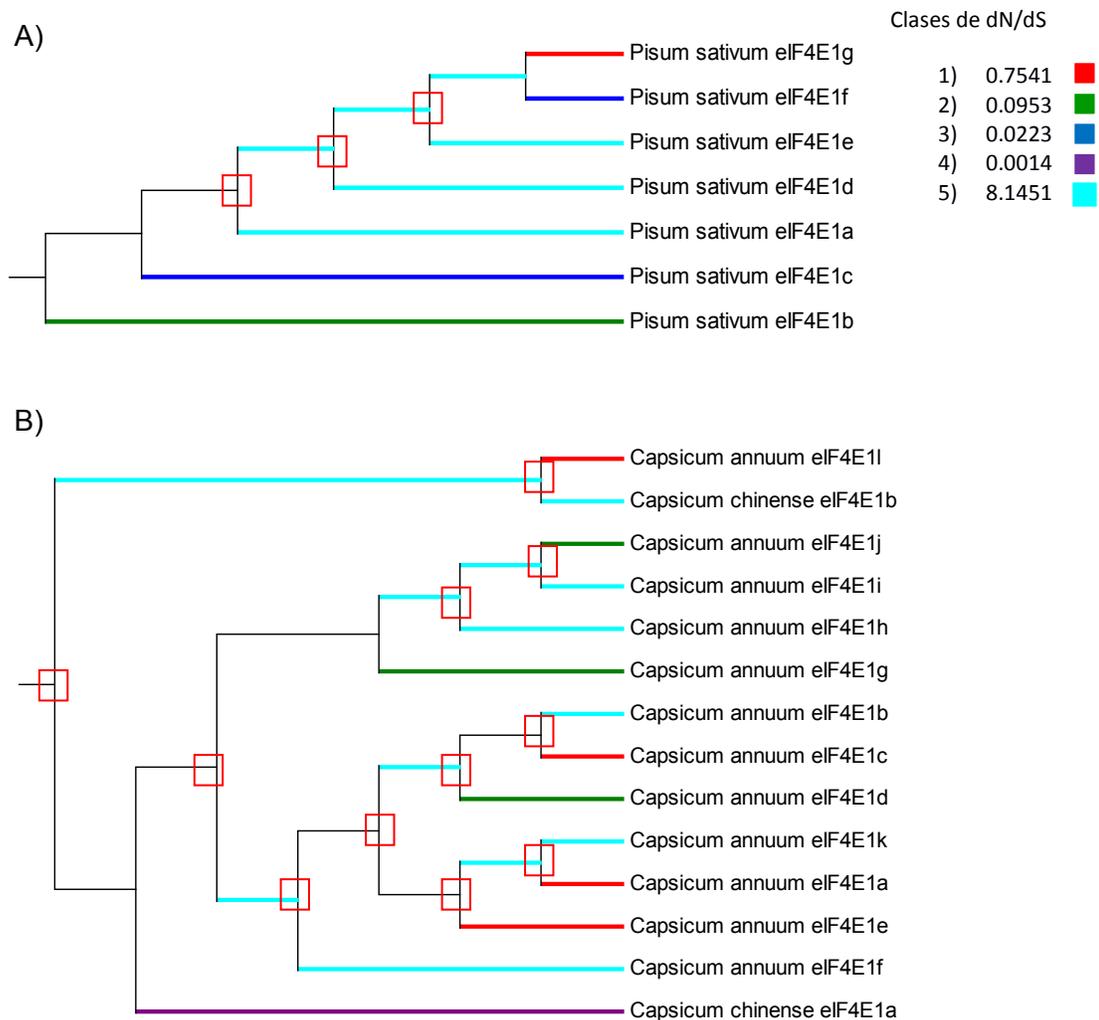


Figura 34. Se muestran dos subárboles tomados de la primera filogenia enraizada, el color de la rama indica a qué clase de dN/dS pertenece. En A) subárbol que agrupa las secuencias de eIF4E de la clase estructural I de *Pisum sativum*; en B) subárbol que agrupa las secuencias de de la clase estructural I de *Capsicum annuum*. El recuadro rojo indica un evento de duplicación del cual uno de los duplicados resultantes presenta selección natural positiva.

Los subárboles de las figuras 32, 33 y 34 presentan ramas bajo selección positiva, e involucran subárboles de distintos linajes orgánicos y distintas clases estructurales; sin embargo, la mayoría de estos que presentan ramas con evidencia de selección natural positiva pertenecen a la clase estructural I con excepción de un solo subárbol de la clase estructural II (figura 33 panel B). También notamos que algunas especies presentan múltiples secuencias parálogas derivadas de rondas de duplicación. La evidencia de presión de selección positiva en estas ramas podría indicarnos que todas las secuencias están aún en el proceso de evolución o en proceso de adquisición de nuevas funciones.

5.4 Detección de Selección Natural en eIF(iso)4E de Viridiplantae.

La detección de selección natural en eIF(iso)4E de Viridiplantae se realizó mediante el método FEL (Kosakovsky-Pond et al 2005b) habilitado para la detección selección natural sitio-rama; los modelos sitio-rama pueden detectar selección positiva que sólo afecte algunos sitios sobre ramas específicas. Como rama *foreground* se tomó la rama de la cual se origina el subárbol de eIF(iso)4E de Viridiplantae, la hipótesis biológica para la elección de esta rama como la rama *foreground* es que se originó de un evento de duplicación y la detección de sitios bajo selección positiva en esta rama pueda explicar el mantenimiento de este duplicado en el genoma. Como rama *background* se tomó el resto del árbol. En la figura 35 se muestra la rama *foreground*, y en la figura 36 se muestra las longitudes de rama del clado de eIF4E de Viridiplantae.

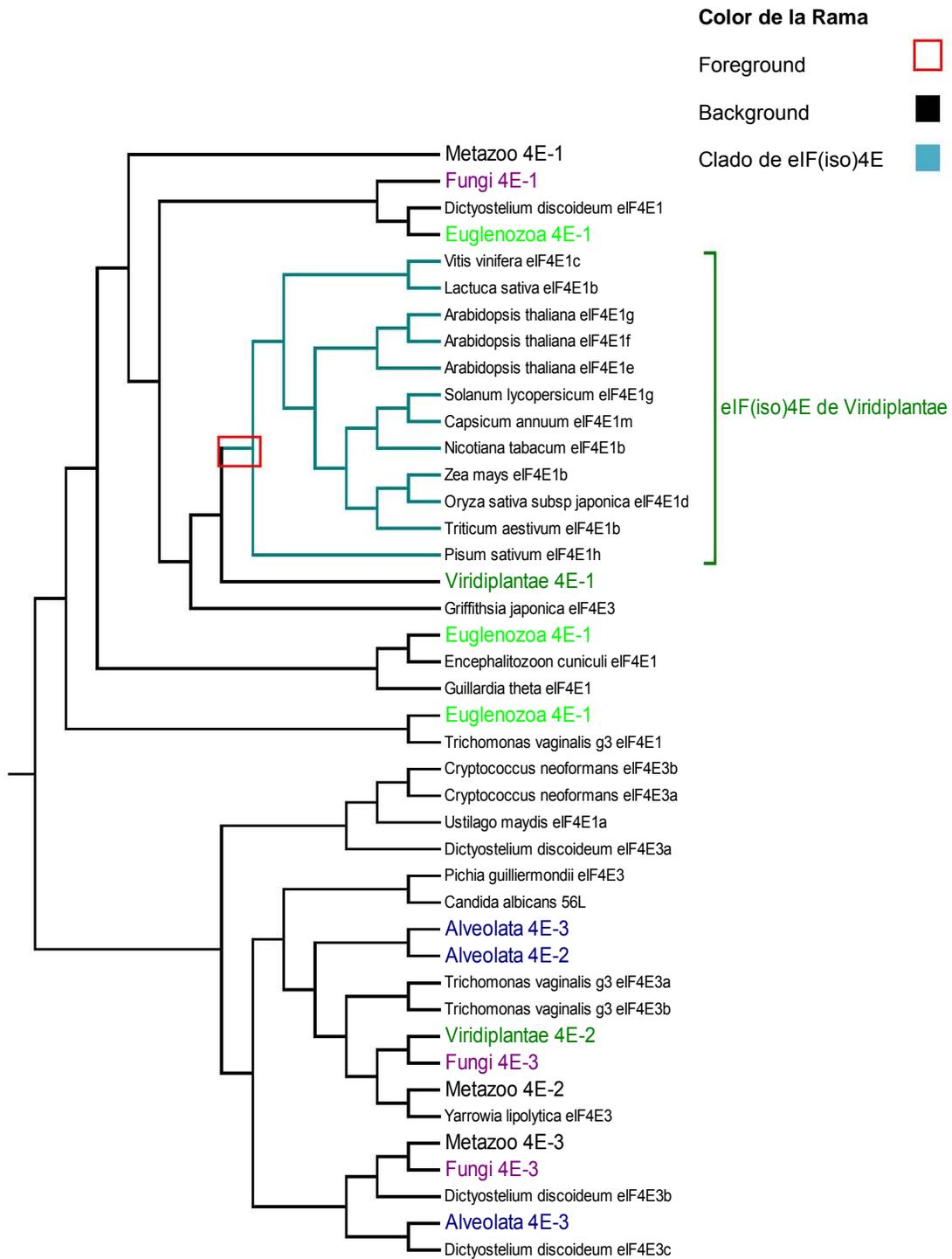


Figura 35. Árbol enraizado de la reconstrucción filogenética, mostrada en la figura 31; el corchete indica el subárbol de eIF(iso)4E de Viridiplantae, las ramas en color azul indican el clado de eIF(iso)4E de Viridiplantae; la rama tomada como rama *foreground* es la que origina al subárbol de las eIF(iso)4E de Viridiplantae y se indica con el recuadro rojo y las ramas *background* se muestran en color negro. Los valores de *bootstrap* no son mostrados.

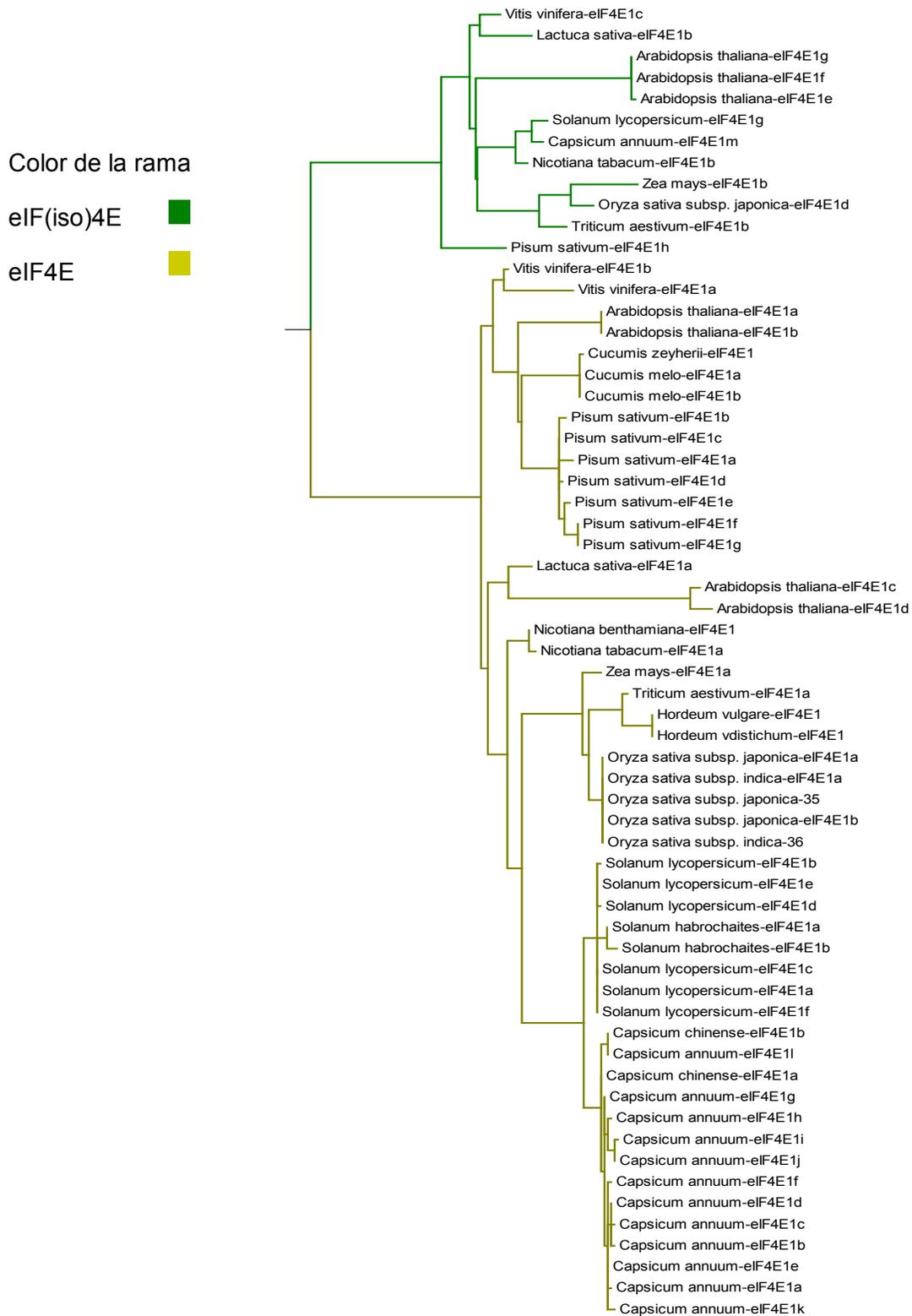


Figura 36. Subárbol de la filogenia enraizada, mostrada en la figura 31, que incluye el clado de eIF4E de Viridiplantae con longitudes de ramas, el color de la rama indica a que variante pertenece, en color verde oscuro eIF(iso)4E y en color verde claro eIF4E. Los valores de *bootstrap* no son mostrados.

En la figura 36 se muestra el subárbol tomado de la filogenia enraizada, que incluye solamente al clado de eIF4E de Viridiplantae, con longitudes de rama, que corresponden al número de cambios entre secuencias. Podemos observar que la longitud de la rama que da origen al clado de la variedad de eIF(iso)4E (en color verde oscuro), es más corta comparada con la rama que origina a el clado de las secuencias de eIF4E (en verde claro), indicando que las secuencias de este ultimo clado han sufrido un mayor número de cambios después del evento de duplicación del cual se originaron.

Son muy pocas las posiciones que presentan evidencia de selección natural positiva, la mayoría presenta régimen selectivo neutral. Dentro de los que están bajo selección positiva su valor de ω es mayor a 1 sin embargo no es un valor muy grande, la posición (en el alineamiento) 257 presentan un valor de ω de 1.33815, la posición 725 presenta un valor de 1.5422 y la posición 788 de 1.3818, lo cual nos está indicando que en estas tres posiciones las sustituciones no sinónimas son más frecuentes en comparación con las sustituciones sinónimas, teniendo como resultado cambios de los aminoácidos para esas posiciones. En las secuencias de eIF(iso)4E estas posiciones presentan un residuo distinto al que se observa en las posiciones homólogas de eIF4E de Viridiplantae. En la tabla 5 se muestra las posiciones del alineamiento, en las cuales se detecto evidencias de selección natural positiva; mientras que en la tabla 6 se muestran que residuo de aminoácido se presenta en eIF4E y eIF(iso)4E en esas posiciones.

Tabla 5. Posiciones que presenta evidencia de selección positiva, indicando la posiciones en el alineamiento y en la secuencia de eIF(iso)4E de *A. thaliana* y de eIF4E de *T. aestivum*, así como el valor de dN/dS en cada posición.

Posición en el alineamiento	Posición en la secuencia de eIF(iso)4E de <i>A. thaliana</i>	Posición en la secuencia de eIF4E de <i>T. aestivum</i>	Valor de dN/dS
257	57	57	1.33815
725	136	149	1.5422
788	149	160	1.3818

Tabla 6. Residuos de aminoácidos que se presentan en eIF4E y eIF(iso)4E en las posiciones detectadas a estar bajo presión de selección positiva.

Posición en el alineamiento	Residuo de aminoácido presente en eIF4E	Residuo de aminoácido presente en eIF(iso)4e
257	Serina	Aspartico o Glutámico
725	Glutámico	Aspartico o Glutámico
788	Lisina	Serina, Arginina, Triptofano o Lisina

En la figura 37 se muestra la distribución de las posiciones mostradas en la tabla 5 en el diagrama de listones de eIF4E de trigo; podemos observar que las tres posiciones se ubican en las asas de la parte cóncava de la estructura. Dos posiciones se ubican en asas que están distantes de la región que interactúa con la estructura *cap*, por lo tanto no estamos seguros si participan directamente en la unión.

A)



B)



Figura 37. Distribución de los sitios detectados a estar bajo selección natural positiva (mostrados en la tabla 5) a través del diagrama de listones 3D de eIF4E de trigo. En color azul se muestran los residuos de Triptófano equivalentes a Trp 56 y 102 de eIF4E de *H. sapiens* y en color rosa los sitios 57, 149 y 160 de *T. aestivum* equivalentes a las posiciones 57, 136 y 149 de *A. thaliana*, bajo Selección Natural Positiva. En A) Vista frontal de la parte cóncava, en B) Vista lateral de eIF4E de trigo.

5.5 Detección de Selección Natural en nCBP de Viridiplantae.

La detección de sitios bajo selección natural positiva se realizó con el método FEL habilitado para la detección sitio-rama en la isoforma nCBP, la rama que se tomó como *foreground* corresponde a la rama que da origen al subárbol que incluye las 5 secuencias de las nCBP de los Viridiplantae. La elección de esta rama para realizar la detección de evidencia de selección natural se debió a que esta isoforma posiblemente lleve a cabo una función diferente a la de eIF4E, y la detección de sitios bajo selección natural positiva nos permitiría inferir que esos sitios posiblemente están relacionados con la nueva función. Como rama *background* se tomó el resto del árbol, como se muestra en la figura 38.

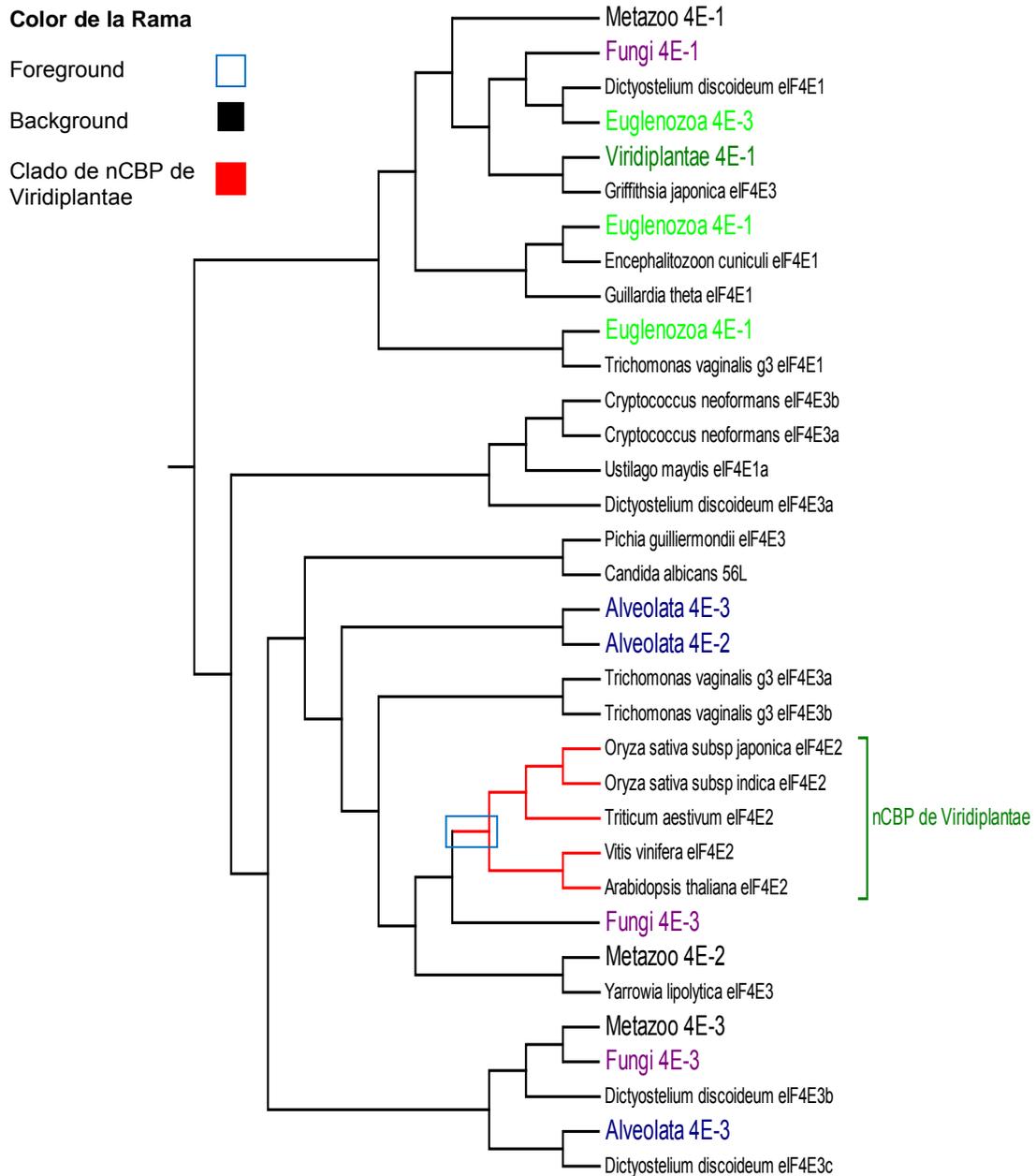


Figura 38. Árbol enraizado de la primera reconstrucción filogenética (figura 31). El corchete indica el clado de nCBP de Viridiplantae, mostrando en color rojo las ramas de las 5 secuencias que componen este clado; el recuadro azul indica la rama que da origen al clado y que fue tomado como rama *foreground*, las ramas *background* se muestran en color negro. Los valores de *bootstrap* no son mostrados.

La prueba de detección de sitios bajo selección natural positiva no identificó sitios bajo presión de selección positiva; posiblemente se deba a que el número de secuencias de la rama *foreground* era muy reducido.

6. Discusión

6.1 Reconstrucción filogenética de la familia de factores de inicio de la traducción eIF4E

Al comparar los árboles que se obtuvieron de las distintas reconstrucciones con el árbol que reportó Joshi *et al.* (2005), a primera vista podemos notar diferencias en la topología. Sin embargo, encontramos la formación de 7 clados que podrían corresponder a 7 de los 8 subgrupos descritos por Joshi *et al.* (2005) y que se muestran en la tabla 7.

Tabla 7. Comparación de los subgrupos obtenidos por Joshi *et al.* (2005) y en las reconstrucciones hechas en este trabajo

Subgrupo (Joshi <i>et al.</i> , 2005)	Clados en las Reconstrucciones Filogenéticas
1) eIF4E-1 de Metazoa, IFE-1, IFE-2, IFE-3 e IFE-5 de Nemátodos	Metazoa 4E-1
2) eIF4E y eIF(iso)4E de plantas	Viridiplantae 4E-1
3) eIF4E-1 de Hongos	Fungi 4E-1
4) eIF4E-2 de Metazoa	Metazoa 4E-2
5) nCBP de plantas	Viridiplantae 4E-2
6) eIF4E-2 de Hongos	Fungi 4E-3
7) eIF4E-3 de Metazoa	Metazoa 4E-3

El subgrupo 8 en nuestras reconstrucciones filogenéticas podría estar formado por distintos clados agrupando secuencias de distintas clases estructurales, que a diferencia del árbol de Joshi (en el cual todas las secuencias de Protistas se encuentran agrupadas en un sólo clado, sin importar a que clase estructural pertenezcan) correspondería a los linajes orgánicos Alveolata, Parabasalide, Euglenozoa, Criptophyta y Mycetozoa. Por otra parte no se observa el subgrupo 6 que correspondería a las secuencias de Fungi pertenecientes a la clase estructural II. Sin embargo, dentro de los diversos clados que observamos de Fungi, que de acuerdo a la clasificación estructural (Joshi *et al.*, 2005) pertenecen a la clase estructural III uno de ellos podrían ocupar el lugar de este subgrupo,

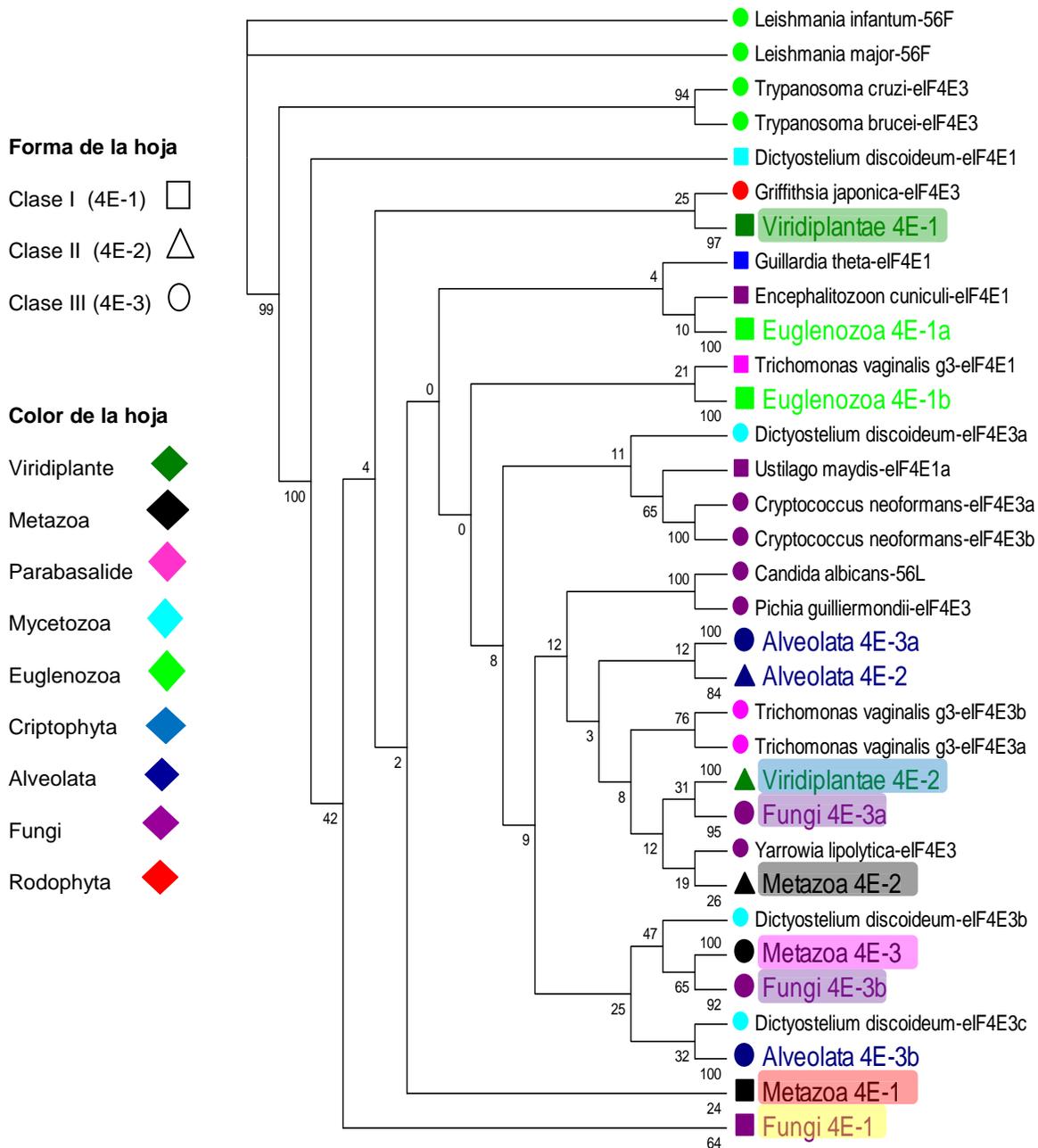


Figura 40. Árbol filogenético sin enraizar, de la región “core” del alineamiento de Pfam con algunas ramas colapsadas. Las ramas colapsadas incluyen varias secuencias de la misma clase estructural y linaje orgánico, y están indicadas por el nombre del linaje orgánico seguido de la clase estructural en un tamaño de letra mayor, y el color correspondiente al color de la hoja. Los recuadros sombreados indican las 7 clados que corresponderían a los 7 subgrupos que observó Joshi *et al.* (2005) en su árbol, el recuadro sombreado en rojo equivale al subgrupo 1, el recuadro sombreado en verde al subgrupo 2, el recuadro sombreado en amarillo al subgrupo 3, el recuadro sombreado en negro al subgrupo 4, el recuadro sombreado en azul al subgrupo 5, el recuadro sombreado en morado al subgrupo 6 y el recuadro sombreado en rosa al subgrupo 7. Los números en los nodos internos indican los valores de *bootstrap*.

Las posibles discordancias entre la topología de los árboles presentados aquí y la obtuvieron Joshi y colaboradores (2005), podría deberse a distintos factores: uno de ellos sería el tipo de datos empleados en ambas reconstrucciones, en la reconstrucción hecha por Joshi y colaboradores (2005) se emplearon como datos secuencia de nucleótidos, mientras que en las reconstrucciones filogenéticas obtenidas en este trabajo, se emplearon secuencias de aminoácidos. Qué tipo de dato se debe de emplear para generar una filogenia o cual es el mejor, es aún un tema de debate. Los caracteres que evolucionan lentamente han sido favorecidos sobre caracteres que evolucionan rápidamente para inferir relaciones filogenéticas profundas, por ello, han sido usados con mayor predilección el carácter aminoácido en lugar del carácter nucleótido cuando se busca resolver filogenias profundas como sería el caso aquí (Simons *et al.*, 2002a); sin embargo, los caracteres que evolucionan lentamente no han demostrado convincentemente su superioridad sobre los caracteres que evolucionan rápidamente (Simons *et al.*, 2002b). El principal argumento del uso de aminoácidos en las reconstrucciones filogenias es que hay mas estados de carácter para aminoácidos (20 estados de carácter para aminoácidos, basados en el código genético universal) que para nucleótidos (4 estados de carácter para nucleótidos) (Harrison y Langdale, 2006), esto incrementa el espacio de estado de carácter lo que hace que la convergencia sea menos probable (Simons *et al.*, 2002b). El argumento en contra del uso de nucleótidos es que con sólo 4 posibles estados de carácter en una posición la presencia de ellos en los nodos terminales puede darse por azar (Harrison y Langdale, 2006). Sin embargo incrementando el número de caracteres de nucleótidos (aumentando la longitud de la secuencia) puede permitir una mejor resolución del árbol (Simons *et al.*, 2002b).

Otro factor que hay que tener en cuenta es la cantidad de secuencias empleadas en cada reconstrucción y en la forma que se obtuvieron; en el trabajo de Joshi *et al.* (2005) las secuencias empleadas se adquirieron desde el GenBank NR y dbEST usando secuencias de nucleótidos codificantes y de aminoácidos de eIF4E-1, eIF4E-2 y eIF4E-3 de *M. musculus*, eIF4E y eIF(iso)4E de *T. aestevium*, nCBP de *A. thaliana*, IFE-1,2,3,4 y 5 de *C. elegans* y eIF4E *S. cerevisiae* como sonda para la búsqueda con BLAST de secuencias homólogas (Joshi *et al.*, 2005). El número de secuencias empleadas para hacer la filogenia no se especifica, solamente se menciona que son más de 100. Las secuencias aquí empleadas para las reconstrucciones se tomaron de la base de datos de Pfam, en la

cual las familias se construyen con base en alineamientos múltiples de secuencias y modelos ocultos de Markov (HMMs). Pfam está compuesta de 2 bases de datos: Pfam-A, en donde las familias están curadas manualmente y Pfam-B en la cual las familias se generan automáticamente. Aunque de menor calidad, las familias de Pfam-B pueden ser empleadas en ciertos análisis. La familia de eIF4E se tomó de la base Pfam-B lo cual puede generar cierto margen de error en la reconstrucción. El número de secuencias empleadas para la primera reconstrucción fue de 263, la inclusión de más secuencias de diversos linajes orgánicos puede influir en obtener una topología distinta a la reportada por Joshi *et al.* (2005), así como una mejor resolución de las relaciones filogenéticas de ciertos clados o la presencia de algunos no encontrados, como el caso de los clados de la clase estructural III de Fungi.

Por otra parte, un factor que influye en gran medida sobre la topología del árbol es con que método fue construido: el árbol de Joshi *et al.* (2005), se construyó mediante el empleo del método de Neighbour-Joining. La principal ventaja de este método es que el tiempo requerido para la reconstrucción es menor en comparación con otros métodos, como máxima verosimilitud y métodos Bayesianos. Sin embargo, tiene las desventajas de que sólo construye un árbol, lo cual no permite examinar las diferentes topologías posibles, y el resultado puede verse afectado por el orden en el que se adicionan las secuencias al árbol inicial (Harrison y Langdale, 2006). Al ser un método basado en distancia evolutiva, la presencia de sustituciones múltiples en un mismo sitio puede no reflejar la distancia verdadera entre secuencias, y hace que artificialmente parezcan cercanas. Mientras el método que empleamos en las reconstrucciones de este trabajo, se obtuvieron con un método que es considerado más robusto, la máxima verosimilitud cuya principal desventaja es el tiempo computacional que requiere; sin embargo, analiza un mayor número de topologías y escoge solamente una, de acuerdo al criterio de optimización. Este método es más robusto ya que es menos sensible a modelos incorrectos y a la atracción de ramas largas (Huelsenbeck y Rannala, 1997). Además, la inclusión de un modelo de sustitución permite caracterizar mejor el proceso de evolución, el árbol obtenido es más confiable.

6.2 Enraizamiento del Árbol Filogenético.

El enraizamiento de los arboles filogenéticos nos permitió observar más claramente las relaciones filogenéticas, algunas de las cuales coinciden con las mostradas en el árbol de Joshi y colaboradores (2005), lo cual nos indica que si hubieran enraizado su árbol posiblemente obtuviera un árbol enraizado parecido a los nuestros. Sin embargo, una de las principales características de nuestras filogenias es la presencia de diversos clados, que de acuerdo a la clasificación estructural pertenecen a la clase III del linaje orgánico de los Fungi y la ausencia de la de la clase II en este mismo grupo, lo cual difiere con la filogenia de Joshi *et al.* (2005). Las secuencias de los clados de la clase estructural III del linaje orgánico de los Fungi pertenecen en su mayoría al subphyllum Pezizomicotina, en él se agrupan especies parásitas de plantas y de animales. La relación filogenética de estos clados de la clase estructural III de Fungi con los clados de la clase estructura II de Viridiplantae y Metazoa o con el clado de la clase estructural III de Metazoa posiblemente reflejen la relación parasito-hospedero; por otra parte las dos secuencias de la clase estructural II de Fungi que Joshi y colaboradores (2005) presentan en su árbol también pertenecen a este mismo subphyllum, lo que nos permite especular que posiblemente estas dos secuencias se clasificaron de acuerdo a su filogenia y no al criterio estructural que ellos emplearon con las demás secuencias.

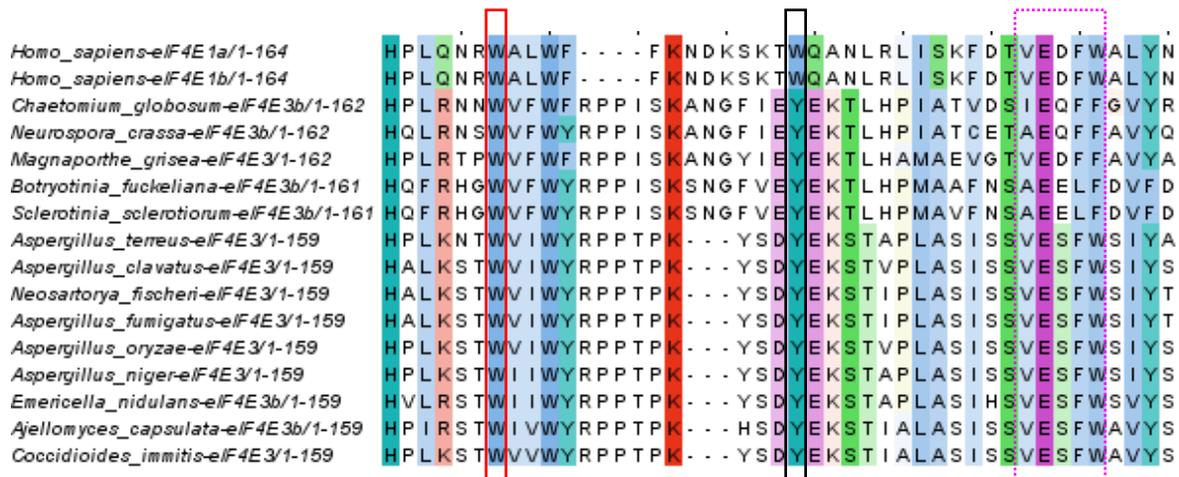


Figura 41. Segmento del alineamiento de la región core de 2 secuencias de *H. sapiens* de la clase estructural I y 14 secuencias de linaje orgánico de los Fungi de la clase estructural III. En el recuadro rojo están indicadas las posiciones homólogas a la posición 43; en los recuadros negros se indican las posiciones homólogas a la posición 56 y en el recuadro punteado en color rosa se indican las posiciones homólogas a la región consenso que interactúa con 4G y 4EBP (Val69XXPhe72Trp73). La numeración es de acuerdo a eIF4E de *H. sapiens*.

En la última filogenia enraizada se observa una distribución característica de los dos clados de la clase estructural III de Fungi; el clado que incluye más secuencias es grupo hermano de la clase estructural II; por otra parte el clado formado por 2 secuencias es grupo hermano de la clase estructural III. Una sola secuencia aparece en el clado que corresponde a la clase estructural II (figura 42).

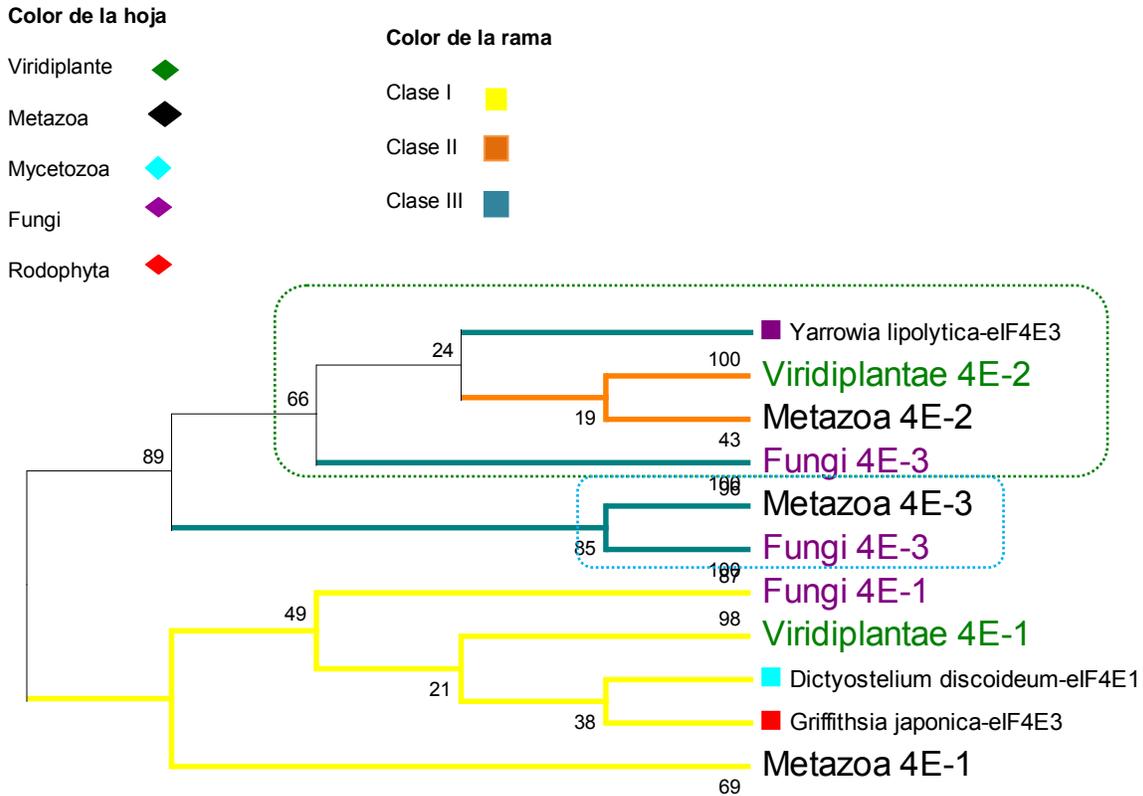


Figura 42. Filogenia enraizada con las ramas colapsadas para una mejor visualización, construida con el alineamiento de secuencias sin sustituciones en las posiciones 46, 69, 72, 73, 113, 130. El color de la rama indica a que clase estructural pertenecen. Las ramas colapsadas están indicadas con el nombre del linaje orgánico seguido de la clase estructural, clase I (4E-1), clase II (4E-2) y clase III (4E-3); las ramas que no están colapsadas están indicadas con el nombre del organismo al cual pertenece la secuencia seguido de la clase estructural al que pertenece, el color de la hoja de estas ramas indica el linaje orgánico al que pertenecen. En los recuadros punteados se indica los clados que corresponden a miembros de la clase estructural III de Fungi. El recuadro verde, indica los subárboles que presentan a los miembros de la clase III de Fungi con la clase II; el recuadro azul indica la relación de grupos hermanos de la clase III de Fungi con la clase III de Metazoo. Los números en los nodos internos representan los valores de *bootstrap*.

Las secuencias de los miembros de la clase estructural III de Fungi, que está como grupo hermano de la clase estructural II, pertenecen a especies de la subclase Eurotiomycetidae

y de los órdenes Eurotiales y Onygenales; mientras que las secuencias más cercanas a la clase estructural III de Metazoa, pertenecen a la subclase Sordariomycetidae, del orden Sordariales. La secuencia que se encuentra en el mismo clado de la clase estructural II, pertenece al orden Saccharomycetales. En el clado que incluye las secuencias de la Subclase Eurotiomycetidae, la división de las secuencias de acuerdo a su orden taxonómico no se da, ya que aparecen en un mismo clado las secuencias de ambos órdenes, cuando deberían de agruparse como clados hermanos (figura 43).

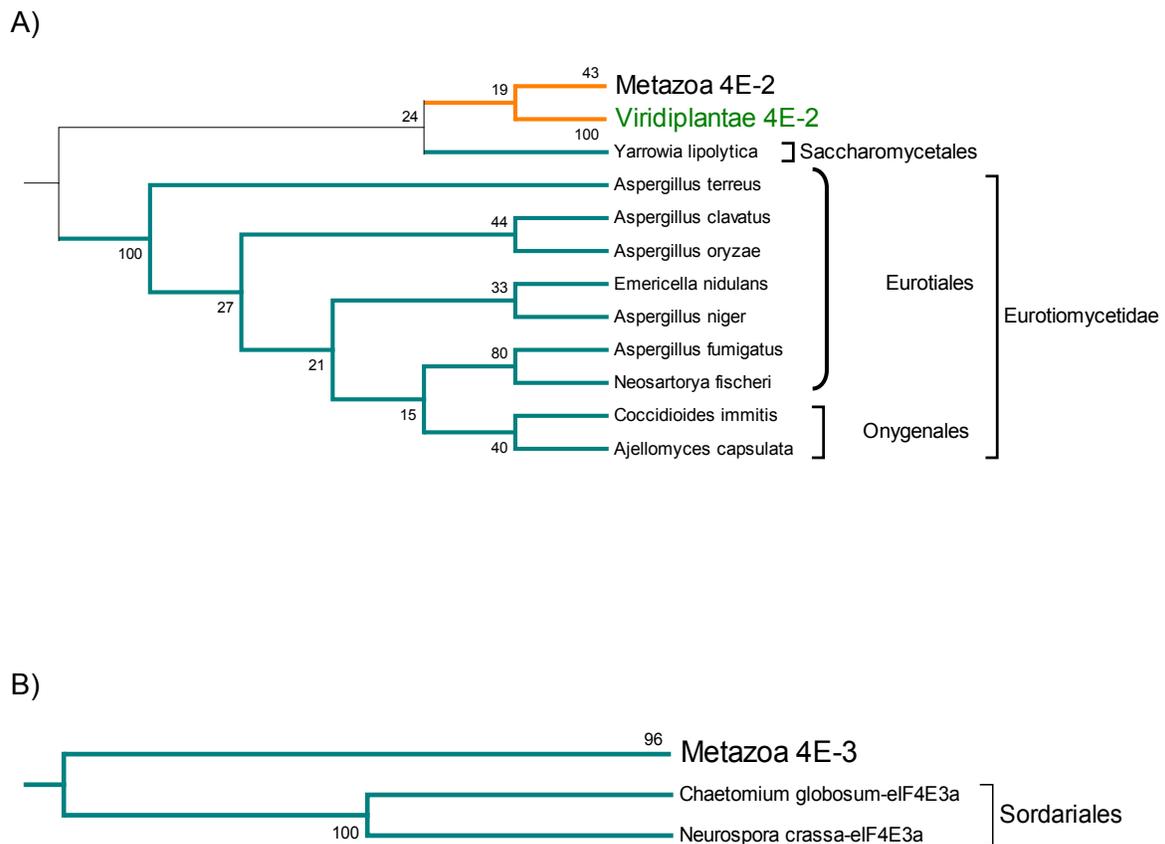


Figura 43. A) Subárbol mostrado en el recuadro punteado de color verde de la figura 39, con la clasificación taxonómica de las especies con secuencias que pertenecen a la clase III de Fungi más cercanas a la clase estructural II de Metazoa y Viridiplantae. B) Subárbol mostrado en el recuadro punteado de color azul de la figura 39, con la clasificación taxonómica de las especies que presentan secuencias de la clase III, en el clado que aparece como grupo hermano del clado de la clase III de Metazoo. Los números en los nodos internos indican los valores de bootstraps.

Mediante el enraizamiento observamos la formación de 2 ramas principales, que tienen como diferencia esencial el residuo que presentan en la posición homóloga a la posición 56 de eIF4E de *H. sapiens* que es uno de los dos residuos que interactúan con el *cap*. Una de las ramas agrupa a las secuencias que presentan un residuo de triptófano en la posición 56 (clase estructural I), mientras que la segunda rama agrupa a las secuencias que presentan un residuo distinto al triptófano como lo sería tirosina, fenilalanina o cisteína (clase estructural II y III). Lo que nos indica que, posiblemente después del evento de duplicación que originó ambas ramas, la segunda rama quedó libre de restricciones funcionales permitiendo sustituciones en la posición 56. La posible acumulación de sustituciones en las secuencias agrupadas en esta rama, quizás les permita desempeñar una función distinta a la canónica de eIF4E (participación en el inicio de la traducción mediante el reconocimiento y unión a la estructura *cap*), que se lleva a cabo en los miembros de la primera rama que presentan los dos residuos de triptófano esenciales (56 y 102 de acuerdo a la numeración de eIF4E de *H. sapiens*) para la interacción con el *cap*. Posteriormente al primer evento de duplicación, se puede observar que la segunda rama sufre un segundo evento de duplicación, del cual se originan dos subramas, una de ellas agrupa a las secuencias de la clase estructural II de Metazoa, Viridiplantae y el clado de la clase estructural III de Fungi; la segunda rama agrupa a las secuencias de la clase estructural III de los Metazoa y los Fungi. La formación de estas dos subramas nos indica que presentan diferencias a nivel de secuencia y que posiblemente estas diferencias les confieran una función distinta a cada rama. Lo que concuerda con las observaciones que miembros de la clase estructural II (eIF4E-2A o 4EHP de *H. sapiens*, eIF4E-2 de *M. musculus*, IFE-4 de *C. elegans*, eIF4E-2 en *D. melanogaster*) pueden interactuar con la estructura *cap* de los RNAm. Sin embargo, son incapaces de interactuar con eIF4G, por lo tanto su participación en el inicio de la traducción es dudosa y quizás un rol regulatorio sea la función que desempeñe (Joshi *et al.*, 2005) como se ha formulado para el homólogo de 4EHP en *Drosophila* d4EHP donde el correcto desarrollo del embrión requiere de una distribución asimétrica de la proteína *caudal*. Esto se logra mediante la regulación a nivel traduccional e involucra a d4EHP en un complejo con otra proteína, *bicoid*. Mientras d4EHP se une a la estructura *cap* 5' del RNAm de *caudal*, también se une a *bicoid* que interactúa con el extremo 3' del RNAm, previniendo que eIF4F se una y la traducción no puede llevarse a cabo. La expresión de *caudal* por ende es dependiente de la concentración local del complejo d4EHP-Bicoid (Rosettani *et al.*, 2007); la excepción dentro de este grupo es nCBP de *Arabidopsis*

thaliana que se unen a la estructura *cap* y puede interactuar con eIF4G y participar en la traducción (Joshi *et al.*, 2005). Con respecto a la función de la segunda subrama que incluye secuencias de la clase III de Metazoa y Fungi se ha mostrado que la interacción con la estructura *cap* es débil, como sucede con eIF4E-3 de *M. musculus* además es incapaz de rescatar el crecimiento de *S. cerevisiae* carente de un gen funcional de eIF4E. La interacción débil de eIF4E-3 de *M. musculus* sugiere que esta proteína puede estar involucrada en el secuestro de eIF4G resultando en la inhibición de la traducción dependiente de *cap* (Joshi *et al.*, 2005).

En la figura 44 se muestra las relaciones filogenéticas de las diversas isoformas de eIF4E basándonos en las reconstrucciones filogenéticas realizadas y las posibles funciones propuestas para cada una de las ramas.

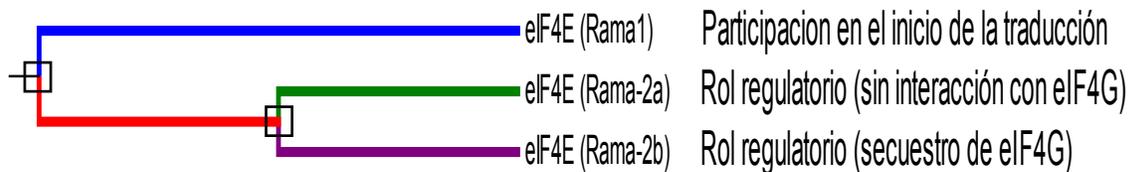


Figura 44. Relaciones filogenéticas de las diversas isoformas con base a las reconstrucciones filogenéticas hechas en este trabajo y las posibles funciones de cada una de las ramas. La rama azul representa a la rama 1 incluye secuencias de la clase estructural I, la rama roja representa la rama 2, la rama verde indica la rama 2a que agrupa secuencias de la clase estructural II de Viridiplantae y Metazoa y secuencias de la clase estructural III de Fungi, la rama morada indica la rama 2b que contiene las secuencias de la clase estructural III de Metazoa y Fungi; los recuadros negros indican los eventos de duplicación.

Basándonos en la figura 44 parecería que una clasificación filogenética de los miembros de esta familia sería más adecuada que la clasificación estructural propuesta por Joshi *et al.*, (2005) actualmente empleada. Algunas posiciones que podrían ser empleadas para diferenciar las secuencias a la rama 1 o 2 se presentan en la tabla 8.

Tabla 8. Posiciones que pueden ser empleadas para diferenciar las secuencias de la rama 1 (clase estructural I) y la rama 2 (clase estructural II y III) la numeración de las posiciones es de acuerdo a eIF4E de *H. sapiens*.

Rama	Posición 56	Posición 116	Posición 162	Posición 167	Posición 180	Posición 198
1	W	Variable	K o R	T	G	Variable
2	Y o C	R o K	Variable	Variable	Variable	K

Las posiciones que pueden ser empleadas para diferenciar las subramas 2a y 2b se muestran en la tabla 9.

Tabla 9. Posiciones que pueden ser empleadas para diferenciar las secuencias de la rama 2a (clase estructural II de Metazoa y Viridiplanate y clase estructural III de Fungi) y la rama 2b (clase estructural III de Metazoa y Fungi) la numeración de las posiciones es de acuerdo a eIF4E de *H. sapiens*.

Rama	Posición 43	Posición 70	Posición 81	Posición 114	Posición 117	Posición 131	Posición 153
2a	W o Y	E	R	I	L o F	E	V
2b	W	Q	Variable	Variable	Variable	K o N	Variable

Mediante las tablas 8 y 9 podría llevarse a cabo la clasificación filogenética de los miembros de la familia de eIF4E.

6.3 Detección de Sitios bajo Selección Natural.

La mayoría de las ramas que están bajo selección natural son ramas terminales mientras que pocas ramas internas y ninguna de las ramas más profundas muestran huellas de selección natural positiva. La ausencia de selección natural positiva en las ramas más profundas, que dan origen a las ramas secundarias 2a y 2b puede deberse a dos posibles escenarios que permiten estos resultados; 1) Errores y falta de poder en los métodos de detección de selección natural positiva y 2) la ausencia de selección natural positiva en la evolución de los genes duplicados.

Respecto al primer escenario, la mayoría de los métodos de detección de selección natural positiva son capaces de detectar sustituciones más recientes con mayor seguridad que sustituciones antiguas, y por ello es difícil detectar selección positiva sobre ramas profundas en una filogenia (Fletcher y Yang, 2010); por otra parte el período de tiempo en el que la selección natural actúa es probablemente muy corto y en algunas ocasiones se puede esperar que sólo actué sobre unos cuantos sitios. Además, el método empleado para la detección de selección natural tiene la ventaja de seleccionar de un conjunto de modelos el que tenga el mejor ajuste estadístico a los datos, a diferencia de otros métodos como los de sitio-rama, donde el requisito de la asignación de un modelo de evolución sobre las ramas es *a priori* y dado un gran número de modelos que pueden ser asignados a las ramas, pueden existir modelos que se ajusten mejor a los datos que el modelo asignado *a priori* (Kosakovsky-Pond y Frost, 2005); sin embargo la detección de selección es a nivel de ramas y no entre sitios, ya que en las pruebas de ramas, la selección positiva es detectada sobre la rama si el promedio de ω sobre todos los sitios es mayor que 1 (Fletcher y Yang, 2010), pero en algunos casos el promedio de ω raramente es mayor que uno debido a que la selección solamente se presenta en un sub-conjunto de aminoácidos específicos (Bielawski y Yang, 2003), y al detectar la selección como un promedio de ω , la señal de estos sitios puede quedar oculta o diluida.

Por otra parte la ausencia de selección natural positiva en las ramas profundas podría deberse a que en el origen de las subramas 2a y 2b no actuó la selección natural adaptativa, por lo tanto “no detectamos selección porque no hubo”. Existen 3 principales razones del porqué un gen duplicado se conserva: 1) la evolución de una nueva función en uno de los duplicados (neo-funcionalización), 2) la división de las funciones ancestrales entre los duplicados resultantes (sub-funcionalización), y 3) la conservación de todas las funciones en ambos duplicados (conservación de genes) (Hahn, 2009). El resultado más importante de la duplicación de genes es el origen de una nueva función, sin embargo una función relacionada más que una función completamente nueva surge después de la duplicación de genes (Zhang, 2003). Como observamos las subramas 2a y 2b (secuencias de las clases estructurales II y III) se originaron de un evento de duplicación y la presencia de sustituciones en posiciones importantes funcionalmente nos permite considerar que las secuencias incluídas en estas ramas se han conservado debido a que presentan una función diferente a la canónica de eIF4E, llevada a cabo por las secuencias

de la rama 1 (clase estructural I), por lo tanto se conservaron por un proceso de neo-funcionalización. Los 2 modelos más comunes para explicar la neo-funcionalización difieren en el papel que juega la selección natural adaptativa en el surgimiento de la nueva función. El modelo Dyhkhuisen-Hartl propone que ninguna de las mutaciones que presenta el gen duplicado es fijada por selección positiva, sino que las mutaciones se acumulan debido a la deriva génica u otros procesos neutrales, y en algún punto del tiempo las condiciones ambientales cambian provocando que la nueva versión del gen duplicado le confiera una ventaja al organismo. La característica importante de este modelo es que ninguna de las mutaciones que surjan recientemente en el gen duplicado aumenta la adecuación con respecto al otro gen, hasta que se fijen y cambien las condiciones ambientales o algún aspecto del *background* genético (Hahn, 2009). Por otra parte, el modelo de adaptación requiere de selección positiva y plantea que después de la duplicación algunas sustituciones neutrales o casi neutrales crean una nueva función en uno de los genes duplicados; sin embargo, esta nueva función no se desempeña correctamente hasta que la selección natural positiva acelera la fijación de mutaciones ventajosas que potencien la nueva función (Zhang, 2003). Hasta que una función llegue a establecerse o se potencie completamente la selección natural positiva cesa y la selección purificadora actúa manteniendo la nueva función. Para genes que codifican proteínas, esto significa que las sustituciones no sinónimas podrían ser acelerada después de la duplicación y posteriormente decaiga para incrementar el efecto de la selección purificadora (Bielawski y Yang, 2003).

De acuerdo a los modelos anteriores y con base en la ausencia de selección natural positiva que arrojan los resultados de detección, nos indica que posiblemente el modelo Dyhkhuisen-Hartl de neo-funcionalización sea el más adecuado para explicar la evolución de los genes duplicados de la clase estructural II y III, de acuerdo a Joshi *et al.* (2005), o 2a y 2b en nuestra filogenia.

6.4 Detección de Selección Natural en eIF(iso)4E de Viridiplantae.

Sólo se detectaron tres posiciones que estarían bajo presión de selección, que se ubican en las asas cercanas a la parte cóncava de eIF4E (figura 37). Sin embargo, no sabemos

si estas posiciones le confieren una función distinta a la isoforma eIF(iso)4E, pero al parecer no son esenciales estructuralmente, ya que no coinciden con las posiciones que se ha propuesto que estarían conservadas en eIF4E ni en las posiciones que se han identificado como características de plantas (linaje-especifico) (Mozingo *et al.*, 2007). Por otra parte, así como han identificado residuos linaje-específicos, también los han identificado para isoformas en plantas (isoforma-especifico), esperaríamos que las posiciones que presentan evidencia de presión de selección positiva coincidieran con algunas de estas posiciones, que son consideradas excelentes candidatos para ser los sitios claves para funciones específicas de isoforma (Mozingo *et al.*, 2007), lamentablemente no fue así, no hubo coincidencia. En la tabla 10 se muestran el residuo característico en cada una de las isoformas.

Tabla 10. Posiciones y residuos específicos de isoformas en plantas, (Mozingo *et al.*, 2007)

Posición del residuo	Residuo en eIF4E	Residuo en eIF(iso)4E
53	Pro	Glu
55	Variable	Lys
59	Gln/Lys	Gly
83	Asn	Asp/Glu
94	Gly	Asp/Ser
99	Cys	Leu
103	Lys	Gly
112	Ile/Val	Glu
127	Lys	Variable
128	Ser	Leu/Phe
134	His/Tyr	Glu
183	Gln	Lys
193	Ser/The	Lys

La mayoría de estos residuos específicos de isoforma son localizados en posiciones relativamente accesibles sobre la superficie de la proteína, de los cuales 5 de esos residuos (53, 55, 59, 103 y 112) son localizados en los loops relativamente cercanos al sitio de unión del m7GTP, sugiriendo que eIF4E y eIF(iso)4E pueden diferir en sus interacciones con el *cap* del RNAm o una región del RNAm cercana a él *cap*. Otros distintos residuos específicos de la isoforma (83, 127, 128, 134, 183 y 193) están también localizados sobre la superficie de la proteína, aunque relativamente lejos del sitio de unión del *cap* (Mozingo *et al.*, 2007).

En la figura 45 se muestra un alineamiento de secuencias en aminoácidos pertenecientes a eIF4E y a eIF(iso)4E de 3 plantas, así como de una secuencia de un mamífero y una levadura, ilustrando las posiciones específicas de eIF4E y las específicas de eIF4E de plantas e isoforma (Mozingo *et al.*, 2007).

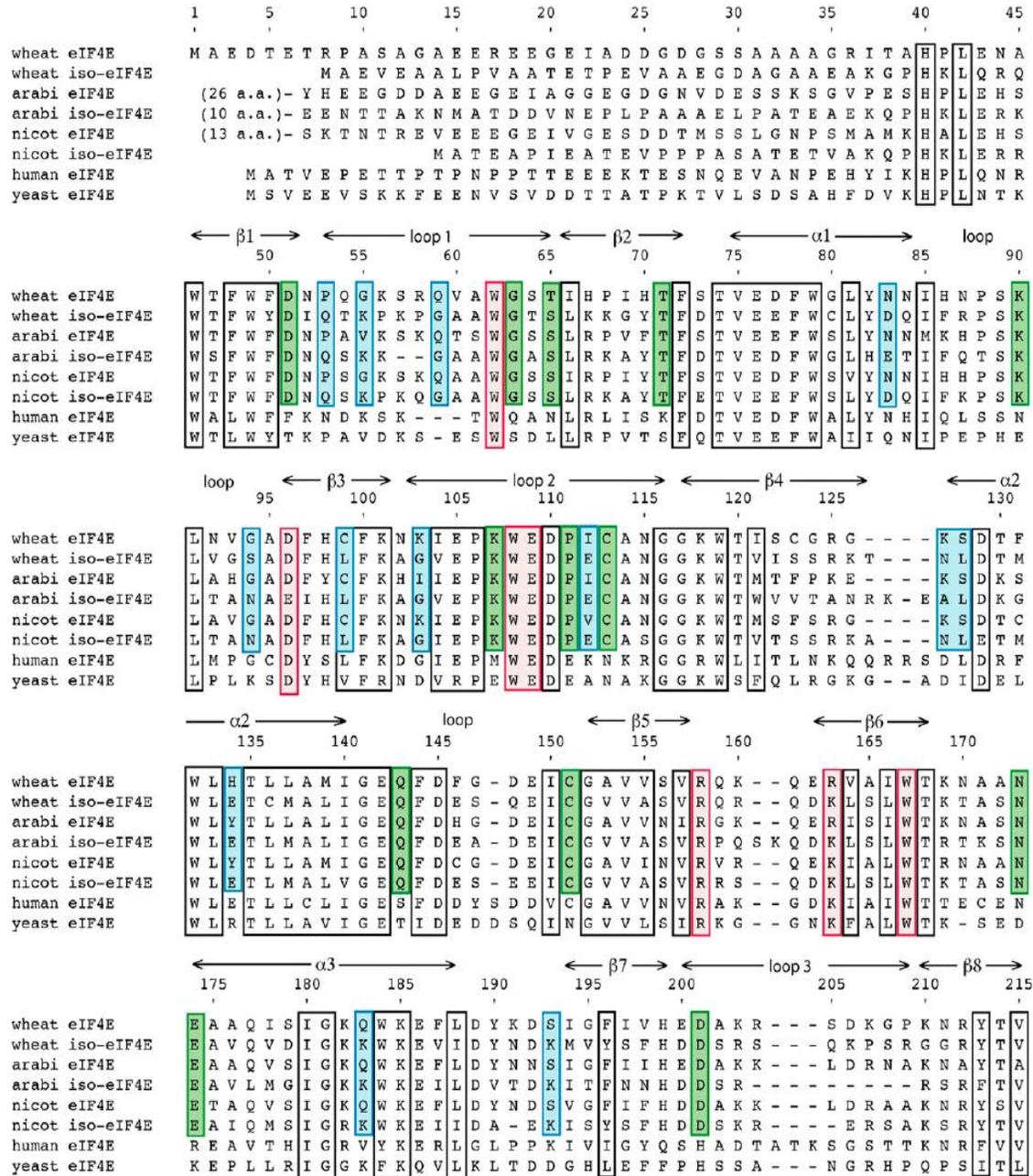


Figura 45. Secuencias de aminoácidos de eIF4E y sus isoformas de plantas superiores alineadas con eIF4E de humano y levadura. Las secuencias son de eIF4E y eIF(iso)4E de Trigo, *Arabidopsis* (arabi), y *Nicotiana tabacum* (nicot), alineadas con eIF4E de *Homo sapiens* (human) y levadura. Los aminoácidos que están involucrados con m7GTP están en recuadros rojos. Los aminoácidos que son importantes para estabilizar la estructura de la proteína están en recuadros negros; aquellos están localizados en el core hidrofóbico o involucrados en otra interacción ternaria y están bien conservados a través de los eucariontes. Los aminoácidos específicos de plantas están en recuadros en verdes; los aminoácidos específicos de plantas incluyen Cys-113 y Cys-151, los cuales forman el puente disulfuro en la estructura del cristal de eIF4E, tal como otros distintos residuos de superficie. Los recuadros azules indican aminoácidos que están más del 90% conservados en una isoforma de eIF4E de planta, mientras que es variables o conservado como otro tipo de residuo en eIF(iso)4E y, por ello, indicando si la proteína pertenece a la familia de eIF4E o eIF(iso)4E. La mayoría de los residuos específicos de la isoforma están relativamente accesibles sobre la superficie de la proteína (Mozingo *et al.*, 2007)

Claramente notamos que ninguna posición que se detectó a estar bajo selección natural positiva coincide con las posiciones específicas de isoforma, de linaje-específicos o importantes estructuralmente de eIF4E. Tampoco sabemos si las posiciones detectadas le confieren una función distinta a la isoforma de eIF(iso)4E de plantas; con lo cual surge la pregunta ¿por qué tener una copia extra, que al parecer no tiene una función distinta? Tómese en cuenta que las secuencias de aminoácidos de eIF4E y eIF(iso)4E son aproximadamente 50% conservada e incluyen los 2 residuos de triptófano encargados del reconocimiento de la estructura *cap* en el RNAm y ambas tienen un número de funciones similares, apoyan la traducción *in vitro*, facilitan la actividad de helicasa dependiente de ATP, y exhiben actividad de ATPasa dependiente de RNA (Miyoshi *et al.*, 2008).

Los mecanismos que explican porque un duplicado se mantiene en el genoma de un organismo son: la neo-funcionalización, la sub-funcionalización y la conservación de genes (Hahn, 2009). Dejando de lado la hipótesis de la neo-funcionalización como la razón de la conservación del gen duplicado (debido a la falta de evidencias), podemos emplear los dos mecanismos restantes para explicar la conservación de los duplicados de eIF4E.

Los dos modelos más comunes para explicar el mecanismo de conservación de genes son el modelo de redundancia y el modelo de cantidad. El primer modelo de redundancia considera que los genes duplicados que mantienen la misma función ancestral se conservan por si las mutaciones inhabilitan uno de los genes duplicados (Hahn, 2009). Por otra parte el modelo de cantidad plantea que ambos genes se conservan porque el producto de estos genes se necesita en grandes cantidades y es una ventaja (Hahn, 2009). Un ejemplo de esto se observa principalmente en genes que son altamente expresados cuyos productos se requieren en grandes cantidades, como lo son genes de rRNAs e histonas (Zhang, 2003).

A menos que la presencia de los dos genes con funciones idénticas sea una ventaja, es improbable que se mantengan en el genoma, se predice que ambos duplicados pueden ser mantenidos cuando ellos difieren en algunos aspectos de sus funciones, lo cual puede

ocurrir mediante sub-funcionalización (Zhang, 2003). La sub-funcionalización se define como la división de las funciones ancestrales entre los duplicados resultantes, la división de estas funciones pueden comprender patrones de expresión (por ejemplo, una expresión diferencial en múltiples tejidos), la función de la proteína (funciones llevadas mediante diferentes sitios activos del mismo péptido) o alguna otra función genética. Existen dos modelos principales para explicar la sub-funcionalización, uno de ellos es el modelo de Duplicación-Degeneración-Complementación que no requiere mutaciones adaptativas para el mantenimiento de los duplicados y plantea que uno de los dos genes resultantes de la duplicación sufre la pérdida de una de las funciones ancestrales por mutaciones, posteriormente el gen duplicado sin mutar sufre una mutación que resulta en la pérdida de una función diferente, los 2 genes duplicados se complementan uno con otro y el organismo requiere ambos genes para una función apropiada. El segundo modelo es conocido como el modelo de Especialización y plantea que hay un conflicto entre las dos funciones llevadas a cabo por el gene ancestral, de forma que no se puede mejorar el desempeño de una función sin afectar la otra. Este conflicto adaptativo no puede ser resuelto hasta que la duplicación de genes permita a uno de los duplicados resultantes especializarse mediante el mejoramiento de una de las dos funciones, gracias a la selección positiva (Hahn, 2009).

Ambos mecanismos podrían ser adecuados para explicar la presencia y conservación de la isoforma de eIF4E; ya que ambas proteínas presentan un número de funciones similares y sobre todo apoyan la traducción *in vitro* (Miyoshi *et al.*, 2008), podríamos especular que la presencia del gen de eIF(iso)4E es debida a la necesidad de altas cantidades de proteína, por lo tanto la presencia de una copia extra del gen proporciona un aumento en la adecuación, como lo menciona el modelo de conservación de genes debido a la cantidad; sin embargo la relativa abundancia y actividad de las isoformas de eIF4E puede variar, por ejemplo se ha demostrado que las isoformas de eIF4E pueden unirse a estructuras cap 5'-hipermetiladas, así como la cantidad de eIF(iso)4E en extractos de germen de trigo es de 3 a 5 veces más alta que la cantidad de eIF4E, y los transcritos de *AteIF4E1* (gen que codifica eIF4E1 en *A. thaliana*) han sido detectados en todos los tejidos de la planta excepto en raíz, mientras que el RNAm de *AteIF(iso)4E* (gen que codifica eIF(iso)4E1 en *A. thaliana*) se detecta en los tejidos jóvenes como los meristemas, semillas jóvenes y flores (Duprat *et al.*, 2002). Estas diferencias en los

patrones de expresión y cantidad podrían ser explicadas mejor mediante el mecanismo de conservación de genes de sub-funcionalización, dentro del cual el modelo de Duplicación-Degeneración-Complementación parece explicar mejor la presencia de la isoforma, ya que no estamos seguros de que el gen ancestral desempeñara más de una función y así apoyar el modelo de especialización.

6.5 Detección de Selección Natural en nCBP de Viridiplantae.

Las pruebas de sito-rama de detección de selección natural positiva para el clado que incluye secuencias de nCBP de Viridiplantae fueron negativas, posiblemente se deba al número reducido de secuencias; el escaso número de secuencias nos permite plantearnos si esta isoforma es característica de todos los miembros de Viridiplantae o de grupos específicos ya que sólo encontramos secuencias de *Vitis vinífera*, *Triticum aestivum*, *Arabidopsis thaliana*, *Oryza sativa* sub especie *indica* y *Oryza sativa* sub especie *japonica*, mientras que el número de especies que presentan eIF4E es mayor.

Lamentablemente estos resultados de ausencia de selección positiva deben de tomarse con reserva, ya que no estamos seguros de que sean correctos debido al número reducido de secuencias empleadas. En el supuesto caso que en verdad no exista selección positiva esto no indica que la isoforma nCBP no presente una función diferente a la canónica de eIF4E ya que de acuerdo a la reconstrucción filogenética el clado de esta isoforma se encuentra en una rama distinta al clado que incluye a eIF4E y eIF(iso)4E, agrupada con el clado de eIF4E-2 de Metazoa (en la rama que denominamos 2a), lo que nos permite plantear que tiene una mayor similitud con estas secuencias, las cuales se consideran que desempeñen una función distinta. En la figura 46 se presenta las relaciones filogenéticas de los 3 miembros de eIF4E que presenta el linaje orgánico de Viridiplantae.

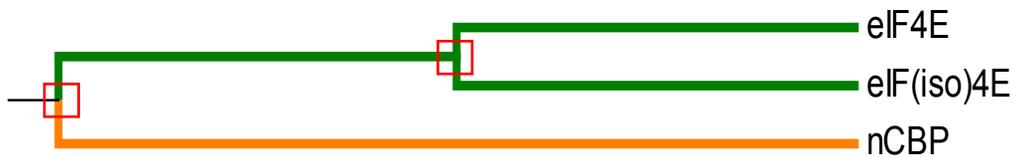


Figura 46. Representación de las relaciones filogenéticas de eIF4E, eIF(iso)4E y nCBP; el color de la rama indica a que clase estructural pertenece, color verde clase estructural I, color naranja clase estructural II, Los recuadros rojos indican eventos de duplicación.

Bajo el supuesto que nCBP presente una función diferente a eIF4E, el mantenimiento de esta isoforma en el genoma podría deberse al mecanismo de neo-funcionalización y la ausencia de selección nos indica que el modelo Dyhkhuisen-Hartl podría ser el más adecuado para explicar el mantenimiento de esta isoforma.

Por otra parte no podemos descartar el mecanismo de sub-funcionalización como el responsable del mantenimiento de este duplicado, ya que a pesar de que es agrupada con eIF4E-2 de Metazoa (los cuales no pueden interactuar con eIF4G a diferencia de nCBP que sí lo puede hacer) puede participar en el inicio de la traducción, y posiblemente un patrón diferente de expresión de este duplicado es lo que lo mantiene en el genoma, y esta expresión diferencial sea el resultado de mutaciones no adaptativas lo cual explica la ausencia de sitios bajo selección positiva, y es esperado de acuerdo al modelo de Duplicación-Degeneración-Complementación.

7. Conclusiones

Inicialmente se asumió que la familia de eIF4E estaba constituida por un sólo tipo de proteína, debido a que sólo un polipéptido de 25 KDa se obtuvo en mamíferos por cromatografía de afinidad. Sin embargo, ahora se reconoce que virtualmente todos los Eucariontes expresan múltiples miembros de la familia eIF4E (Rhoads, 2009). Al parecer la historia evolutiva de la familia de factores de inicio de la traducción 4E (eIF4E) es compleja. Como se observó en la reconstrucción filogenética los eventos de duplicación que dieron origen a los diversos miembros de esta familia son considerables. La duplicación de genes ha sido tradicionalmente considerada como la mayor fuente para la evolución de proteínas con nuevas funciones, lo que permite considerar que los miembros de esta familia desempeñen una función distinta a la original.

La inclusión de más secuencias de los miembros de esta familia y el empleo de métodos más robustos para la reconstrucción filogenética, nos permiten tener una imagen más completa sobre el origen y la evolución de esta familia. La topología del árbol filogenético enraizado difiere con la reconstrucción filogenética obtenida por Joshi *et al.*, (2005) quienes emplearon un menor número de secuencias y un método distinto; adicionalmente se identificaron secuencias pertenecientes al linaje de los Fungi que, de acuerdo al criterio de clasificación estructural (Joshi *et al.*, 2005) pertenecen a la clase III, así como la ausencia de miembros de la clase II en este mismo linaje organísmico, algo que no se ha reportado anteriormente y contrasta con los resultados obtenidos por Joshi *et al.*, (2005). Esto permite cuestionar el actual criterio estructural de clasificación para los miembros de esta familia; al parecer una clasificación basada en las relaciones filogenéticas de los distintos miembros de esta familia podría ser más adecuada. Tomando esto en cuenta se identificaron algunas posiciones que podrían ser consideradas como diagnósticas para realizar esta clasificación.

Por otra parte, los métodos de detección de selección natural nos ayudan a dilucidar si este mecanismo estuvo involucrado en la evolución de esta familia. La ausencia de selección natural positiva en el origen de las clases estructurales II y III, permite plantear que los mecanismos que actuaron sobre los duplicados fueron la deriva génica o algunos

otros procesos neutrales (modelo Dykhuisen-Hartl), bajo el supuesto de que ambas clases presentan una función distinta a la original (neo-funcionalización).

Como un caso particular tenemos la evolución de los miembros de la familia de eIF4E pertenecientes al linaje de Viridiplantae. Este linaje sólo presenta secuencias de las clases estructurales I y II. Dentro de las secuencias que pertenecen a la clase estructural I encontramos 2 isoformas eIF4E y eIF(iso)4E, la conservación del duplicado podría deberse a la sub-funcionalización mediante el modelo de Duplicación-Degeneración-Complementación, debido a la ausencia de selección natural positiva. Con respecto a la conservación de la secuencia perteneciente a la clase estructural II, la nCBP, existen dos escenarios diferentes, en el primero planteamos que este duplicado se ha mantenido por que presenta una función distinta (neo-funcionalización) a la de eIF4E y la ausencia de selección positiva estaría explicada mediante el modelo Dykhuisen-Hartl; en el segundo escenario consideramos que la conservación es debida a una expresión diferencial del duplicado (sub-funcionalización).

La falta de información acerca de los roles fisiológicos de los diversos miembros de la familia de eIF4E dificulta establecer el mecanismo exacto de evolución de cada miembro perteneciente a la Familia de eIF4E; sin embargo el trabajo realizado es una buena aproximación al ser un estudio global que incluye un mayor número de secuencias y diversos linajes en comparación con estudios previos, permitiendo la obtención de información más general sobre la evolución de esta familia. Finalmente el empleo de las reconstrucciones filogenéticas complementado con los métodos de detección de selección natural permite obtener una imagen más completa sobre la historia evolutiva de esta familia.

8. Perspectivas

Los RNAm de los eucariontes actuales tiene tres grandes diferencias comparados con los de procariontes, 1) Son monocistronicos, 2) son poliadenilados en el extremo 3', y 3) presentan una estructura cap 5', siendo este último, de gran importancia para el inicio de la traducción dependiente de cap, el mecanismo mediante el cual la mayoría de los RNAm eucarionticos son traducidos.

El mecanismo de la traducción del RNAm de la mayoría de los organismos comienza con un paso fundamental y característico, el cual es el reclutamiento de la subunidad menor ribosomal al RNAm; el cómo se lleva a cabo este paso ha variado en cada uno de los dominios celulares, comenzando desde mecanismo relativamente sencillos, como lo sería la interacción directa del ribosoma con el RNAm mediante secuencias de reconocimiento y el empleo de unos pocos factores proteicos de inicio de la traducción, hasta mecanismo más elaborados que requieren un gran número de distintos factores de inicio de la traducción, como en el caso de los eucariontes. Sin embargo, a pesar de la gran cantidad de información, referente al inicio de la traducción dependiente de cap, el cómo cada uno de los distintos factores surgieron y evolucionaron permitiendo el establecimiento de este mecanismo, no se ha resuelto totalmente. Dentro de los distintos factores que participan en el inicio de la traducción dependiente de cap, eIF4E sobresale, debido a su relevancia funcional, ya que reconoce y se une a la estructura cap, y por ello también es un blanco de regulación de inicio de la traducción dependiente de cap.

Dentro de este marco, resulta conveniente estudiar el origen del factor de inicio de la traducción eIF4E, debido a que su origen de este factor está altamente relacionado con el origen y el establecimiento del mecanismo de la traducción dependiente de cap. Recordemos que la mayoría de las proteínas son el resultado de la duplicación de proteínas ancestral y que posteriormente divergieron adquiriendo una función diferente a la función original; por otra parte algunos mecanismo celulares se originaron mediante la inclusión de proteínas que participaban en un mecanismo diferente, realizando la misma función, como en el caso de algunas rutas biosintéticas. Teniendo esto en cuenta se

podría realizar una búsqueda de homología remota a eIF4E para saber si esta participaba en un mecanismo distinto al del inicio de la traducción dependiente de cap y posteriormente fue reclutada para participar en este mecanismo. La búsqueda de homología podría llevarse a cabo mediante el programa PSI-BLAST tomando como sonda la secuencia de eIF4E; de las proteínas que sean identificadas homólogas a eIF4E nos interesaría aquellas que realicen una función diferente y que no pertenezcan a la familia de eIF4E, ya que nos indicarían que se originaron de la misma secuencia ancestral.

9. Bibliografía.

Aagaard J.E. y Phillips P. (2005). Accuracy and power of the likelihood ratio test for comparing evolutionary rates among genes. *J. Mol. Evol.*, 60:426-433

Abascal, F., Zardoya, R. y Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, 21:2104-2105.

Adachi, J. y Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.*, 42:459-468.

Bielawski, J.P. y Yang, Z. (2003). Maxima likelihood methods for detecting adaptive evolution after gene duplication. *Journal of Structural and Functional Genomics*, 3:201-212.

Combe, J.P., Petracek, M.E., Van Eldik, G., Meulewaeter, F. y Twell, D. (2005). Translation initiation factors eIF4E y eIFiso4E are required for polysome formation and regulate plant growth in tobacco. *Plant Mol. Biol.*, 57: 749-760.

Dayhoff, M.O., Schwartz, R.M. y Orcutt, B.C. (1978). A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Washington, DC, pp. 345-352.

Duprat, A., Caranta, C., Revers, F., Menand, B., Browning, K.S. y Robaglia, C. (2002). The Arabidopsis eukaryotic initiation factor (iso)4E is dispensable for plant growth but required for susceptibility to potyviruses. *The Plant Journal*, 32:927-934.

Fletcher, W. y Yang, Z. (2010). The effect of insertions, deletions and alignment errors on the branch-site test of positive selection. *Mol. Biol. Evol.*

Foster, P.G. (2001). The idiot's guide to the zen of likelihood in a nutshell in seven days for dummies, unleashed. A gentle introduction, for those of us who are small of brain, to the calculation of the likelihood of molecular sequences. 1-8.

Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on simple model of sequence data. *Mol. Biol. Evol.*, 14(7):685-695

Goldman, N. y Yang, Z. (1994). A codón-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11(5):725-736.

Goodfellow, I.G. y Roberts, L.O. (2008). Eukaryotic initiation factor 4E. *IJBCB*, 40:2675-2680.

Guindon, S. y Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, 52(5):696-704

- Hahn, M.W. (2009). Distinguishing among evolutionary models for the maintenance of gene duplicates. *Journal of Heredity*, **100**(5):605-617
- Harrison, C.J. y Langdale, J.A. (2006). A step by step guide to phylogeny reconstruction. *The Plant Journal*, **45**:561-572
- Hernandez, G. y Vazquez-Pianzola, P.(2005). Functional diversity of the eukaryotic translation initiation factors belonging to eIF4 families. *Mechanism of Development*, **122**: 865-876.
- Higgs, P.G. y Attwood, T.K. (2005). Bioinformatics and molecular evolution. Blackwell Publishing.
- Holder, H., y Lewis, P.P. (2003). Phylogeny estimation: traditional and bayesian approaches. *Genetics*, **4**:275-284
- Huelsenbeck, J.P. y Rannala, B. (1997). Phylogenetic methods come of age: testing hypotheses in a evolutionary context. *Science*, **276**
- Huelsenbeck, J.P. y Crandall, K.A. (1997). Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu. Rev. Ecol. Syst*, **28**:437-66
- Jankowska-Anyszka, M., Lamphear, B.J., Aamodt, E.J., Harrington, T., Darzynkiewicz, E., Stolarski, R. y Rhoads, R. E. (1998). Multiple isoforms of eukaryotic protein synthesis initiation factor 4E in *Caenorhabditis elegans* can distinguish between mono- and trimethylated mRNA cap structures. *J. Biol. Chem.*, **273**:10538-10542.
- Jones, D.T., Taylor, W.R. y Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comp. Appl. Biosci.*, **8**:275-282.
- Joshi, B., Cameron, A. y Jagus, R. (2004). Characterization of mammalian eIF4E-family members. *Eur. J. Biochem.* **271**:2189-2203.
- Joshi, B., Lee, K., Maeder, D.L. y Jagus R. (2005). Phylogenetic analysis of eIF4E-family members. *BMC Evol. Biol.*, **5**:48.
- Keane, T.M., Creevey, C.J., Pentony, M.M., Naughton, J.T., y McInerney, J.O. (2006). Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evolutionary Biology*, **6**: 1-17
- Keith, J.M. (2008). Bioinformatics, volumen I: Data, sequences, analysis, and evolution. Springer Science and Bussiness Media, Totowa, NJ.
- Kosakovsky-Pond, S.L., Frost, S.D.W. y Muse S.V. (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**(5): 676-679
- Kosakovsky-Pond, S.L. y Frost, S.D.W. (2005a). A genetic algorithm approach to detecting lineage-specific variation in selection pressure. *Mol Biol Evol*, **22**(3): 478-485

Kosakovsky-Pond, S.L. y Frost, S.D.W. (2005b). Not so different after all: a comparison of methods for detecting amino acid sites under selection molecular biology and evolution **22**(5): 1208-1222

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. y Higgins, D.G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**:2947-2948.

Li, L. y Wang, C.C. (2005). Identification in the ancient protist *Giardia lamblia* of two eukaryotic translation initiation factor 4E homologues with distinctive functions. *Eukaryotic Cell*, **4**, 948-959.

Maddison, W y Maddison, D. (2007) //http:mesquiteproject.org

Massingham T. y Goldman N. (2004). Detecting amino acid sites under positive selection and purifying selection. *Genetics*, **169**:1753-1762.

McKendrick, L., Pain, V.M. y Morley, S.J.(1999). Translation initiation factor 4E. *IJBCB*. **31**:31-35.

Miyoshi, H., Okade, H., Muto, S., Suehiro, N., Nakashima, H., Tomoo, K. y Natsuaki, T. (2008). Turnip mosaic virus VPg interacts with *Arabidopsis thaliana* eIF(iso)4E and inhibits in vitro translation. *Biochimie*, **90**:1427-1434

Monzingo, A.F., Dhaliwal, S., Dutt-Chaudhuri, A., Lyon, A., Sadow, J.H., Hoffman, D.W., Robertus, J.D. y Browning, K.S. (2007). The Structure of eukaryotic translation initiation factor-4E from wheat reveals a novel disulfide bond 1[OA]. *Plant Physiology*, **143**:1504-1518.

Muse, S. V., y B. S. Gaut (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* **11**:715–724.

Nei, M. y Kumar, S. (2000). Molecular evolution and phylogenetics. Oxford University Press.

Ouyang, Z. y Liang, J. (2007). Detecting positively selected sites from amino acid sequences: an implicit codon model. 29th Annual International Conference of the IEEE EMBS.

Page, R.D.M. y Holmes, E.C. (1998). Molecular evolution: a phylogenetic approach. Blackwell Publishing.

Page, R. D. M. (2000). Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny. *Molecular Phylogenetics and Evolution*, **14**:89-106

- Ptushkina, M., Berthelot, K., Von der Haar, T., Geffers, L., Warwicker, J. y McCarthy, J.E.G. (2001). A second eIF4E protein in *Schizosaccharomyces pombe* has distinct eIF4G-binding properties. *Nucleic Acids Res.*, **29**:4561-4569.
- Ptushkina, M., Malys, N. y McCarthy, J.E.G. (2004). eIF4E isoform 2 in *Schizosaccharomyces pombe* is a novel stress-response factor. *EMBO reports*, **5**:311-316.
- Robaglia C. y Caranta C. (2006). Translation initiation factors: a weak link in plant RNA virus infection. *TRENDS in Plant Science*, **11**(1):40-45
- Robalino, J., Joshi, B., Fahrenkrug, S.C. y Jagus, R. (2004). Two zebrafish eIF4E family members are differentially expressed and functionally divergent. *J. Biol. Chem.*, **279**:10532-10541.
- Rhoads, R.E. (2009). eIF4E: New family members, new binding partners, new roles. *The Journal of Biological Chemistry*, **284**(25):16711-16715.
- Rosettani, P., Knapp, S., Vismara, M.G., Rusconi, L. y Cameron, A.D. (2007). Structures of the human eIF4E homologous protein, h4EHP, in its m7GTP-bound and unliganded forms. *J. Mol. Biol.*, **368**:691-705.
- Saitou, N. y Nei, M. (1987). The neighbor-joining method: a new method for reconstruction of phylogenetic trees. *Mol Biol Evol* **4**:406–25
- Salemi, M. y Vandamme, A.M. (2003). *The phylogenetic handbook: a practical approach to DNA and protein phylogeny*. Cambridge University Press.
- Simmons, M.P., Ochoterena, H. y Freudenstein, J.V. (2002a). Conflict between amino acid and nucleotide characters. *Cladistics*, **18**:200-206.
- Simmons, M.P., Ochoterena, H y Freudenstein, J.V. (2002b). Amino acid vs. nucleotide characters: challenging preconceived notions. *Molecular Phylogenetics and Evolution*, **24**:78-90.
- Takahashi, K., y Nei, M. (2000). Efficiencies of fast algorithms of phylogenetic inference under the criteria de maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. *Mol. Biol. Evol*, **17**:1251-1258.
- Vernot, B., Stolzer, M., Goldman, A. y Durand, D. (2007). Reconciliation with non-binary species trees. *Comput. Syst. Biol.*, **6**:441-452
- Von der Haar, T., Ball, P.D. y MacCarthy, J.E.G. (2000). Stabilization of eukaryotic initiation factor 4E binding to the mRNA 5'-cap by domains of eIF4G. *J. Biol. Chem.*, **275**:30551-30555.
- Wang, D.P., Wan, H.L., Zhang, S. y Yu, J.(2009). gamma-MYN: a new algorithm for estimating Ka and Ks with consideration of variable substitution rates. *Biology Direct*. **4**:20

Waterhouse, A.M., Procter, J.B., Martin, D.M.A, Clamp, M. y Barton, G. J. (2009) Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**(9):1189-1191

Whelan, S. y Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol.Biol. Evol.*, **18**: 691–699.

Yang, Z y Nielsen, R. (2000). Estimating synonymous and nosynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* **17**(1):32-43.

Yang, Z. y Nielsen R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites specific lineages. *Mol. Biol. Evol.* **19**(6): 908-917.

Yang, Z. (2006). Computational molecular evolution. Oxford University Press

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol.*, **24**(8): 1586-1591

Zhang, J. (2003). Evolution by gene duplication: an update. *TRENDS in Ecology and Evolution*, **18**(6):292-298.

Zhang J., Nielsen R. y Yang A. (2005). Evaluation of a improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol*, **22**(12):2472-2479.