UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Distribución filogenética de las proteínas de la biosíntesis de carotenoides en bacterias

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

Biólogo

P R E S E N T A:

Oscar Francisco González Gutiérrez

DIRECTOR DE TESIS:
Dr. Luis David Alcaraz Peraza
Ciudad de México, México 2016

Índice

Índice de tablas

Índice de figuras

Índice de Anexos

Resumen

Los carotenoides son compuestos isoprenoides producidos por hongos, plantas, algas y bact

En el caso de las bacterias, los carotenoides actúan principalmente como pigmentos y agentes antioxidantes. Las proteínas que participan en la biosíntesis de carotenoides (BC) se relacionan con las propiedades funcionales y la diversidad de los compuestos carotenoides. Aunque dichas proteínas han sido ampliamente estudiadas en bacterias fotosintéticas, aún se desconoce su presencia e importancia en gran parte de las bacterias no fotosintéticas.

Con el objetivo de describir la biosíntesis de carotenoides (BC) en el dominio Bacteria, se analizó la abundancia y la distribución de las principales proteínas de la BC en los phyla más estudiados de bacterias. Para ello, se buscaron homólogos de 25 perfiles de proteínas (Modelos Ocultos de Márkov o *Hidden Márkov Model*) de la BC, en 149 proteomas de bacterias. En total, se hallaron 253 proteínas homólogas de la BC; de las cuales, las proteínas que participan en las reacciones iniciales de la BC fueron las más abundantes y por lo tanto las más conservadas. Además, se observó que el phylum Cyanobacteria presentó la mayor abundancia relativa de proteínas homólogas de la BC, respecto al resto de los phyla analizados. En este sentido, se observó que la proporción de genes que codifican a las proteínas de la BC es mínima respecto al resto de las proteínas predichas en los proteomas. Con los resultados obtenidos, se pudo observar que la distribución y la abundancia de las proteínas de la BC es heterogénea en el dominio Bacteria. Incluso, se pudieron reconocer patrones en los cuales la presencia de dichas proteínas podría brindar ventajas adaptativas en los diversos hábitos de vida estudiados (*ej*. fotosintético, patógeno, simbiótico y de vida libre).

Introducción

El genoma de bacterias

El genoma es la información genética total de un organismo (Brown, 2002). En procariontes, eucariontes y algunos virus, el genoma está constituido por ácido desoxirribonucleico (DNA) (Coffin, Hughes, & Varmus, 1997), sin embargo, en determinados virus (*ej.* ribovirus) el genoma puede estar formado por ácido ribonucleico (RNA) (Malpica et al., 2002).

Gran parte de la información genética del genoma se encuentra almacenada en el DNA mediante un código genético (Nelson & Cox, 2008a). Dicho código consiste en tripletes de nucleótidos de DNA que pueden ser transcritos a RNA mensajero (mRNA), por la RNA polimerasa (Nelson & Cox, 2008b). Posterior a la formación de mRNA, se lleva a cabo la síntesis de aminoácidos, por los ribosomas. En el mRNA, los tripletes de nucleótidos son conocidos como codones, los cuales indican al ribosoma el inicio de la traducción, los aminoácidos que se incorporarán y el término de la traducción (Lodish et al., 2000). Los codones se encuentran organizados en marcos de lectura. Un marco contiene los codones necesarios para la síntesis de una proteína, incluyendo el codón de inicio y los codones de paro de la traducción.

La expresión de los genes y la codificación de proteínas presentes en el genoma, son importantes debido a que permiten el desarrollo y el mantenimiento de los organismos (Crick, 1956). Cabe mencionar que en el genoma, pueden encontrarse regiones codificantes de proteínas y otras regiones no codificantes, que pueden llevar a cabo otras funciones (*ej.* regulación de la transcripción) (Brown, 2002; Westhof, 2010).

En bacterias, el genoma es contenido comúnmente en un solo cromosoma circular, sin embargo,

también existen cromosomas lineales como en *Borrelia burgdorferi* B31 (Fraser et al., 1997). El cromosoma bacteriano es diferente al que presentan las arqueas y los eucariontes. En bacterias, los cromosomas son compactados por proteínas como la HU y el DNA súperenrollado puede formar bucles (Kleppe, Steinar, & Lossius, 1979). En las arqueas, la proteína HU no está presente y es sustituida por proteínas similares a las histonas de los eucariontes (Griswold, 2008). En eucariontes, los cromosomas se localizan dentro del núcleo y el DNA se encuentra compactado por proteínas como las histonas (Brown, 2002; Devlin, 2004; Griswold, 2008).

Los genomas de bacterias y eucariontes son diferentes en otros aspectos. En bacterias, existen genes extra-cromosómicos que forman parte del genoma pero que se alojan en moléculas conocidas como plásmidos (López-López, López-Gutiérrez, Sainz-Espuñes, & Rosales-Torres, 2005). Un plásmido es una molécula circular o lineal de DNA, independiente al cromosoma, que mantiene características importantes adicionales al genoma cromosómico principal (*ej.* resistencia a antibióticos) (Brown, 2002). Otra diferencia más entre el genoma de los eucariontes y bacterias es el tamaño. El tamaño del genoma en bacterias (< 15 Mb) suele ser menor que el tamaño del genoma de los eucariontes (< 150 000 Mb) (Han et al., 2013; Pellicer, Fay, & Leitch, 2010). Por otra parte, los genomas de bacterias generalmente tienen sólo una copia de cada gene, no presentan intrones y las secuencias repetidas son poco frecuentes (Griswold, 2008). Los intrones son secuencias localizadas entre exones que son eliminadas del RNAm antes de ser traducid, como sucede en eucariontes (Land et al., 2015). Los operones son otra diferencia entre los genomas de procariontes y eucariontes. En procariontes, los operones fueron descritos como un conjunto de genes que se encuentran bajo el control de una sola señal de regulación o promotor (Blumenthal, 2004; Griswold, 2008; Jacob & Monod, 1961). Los operones en bacterias favorecen una transcripción policistrónica por el agrupamiento de varios genes en un solo operón. En eucariontes, comúnmente los genes se transcriben

de manera individual, sin embargo, se ha llegado a considerar la existencia de operones en algunos eucariontes (ej. nematodos, cordados primitivos, platelmintos y trypanosomas) debido a la forma en la que los productos de múltiples genes son producidos a partir de un simple promotor y demás evidencias experimentales (Blumenthal, 2004).

Genómica comparativa

La genómica es la disciplina dedicada a la secuenciación y análisis de los genomas (López-López et al., 2005). A diferencia de la genética que se dedica al estudio individual de genes, la genómica se dedica al estudio completo de los genes en un organismo (Land et al., 2015).

La genómica se divide en genómica estructural, funcional y comparada (Dorcas & Orengo, 2013). La genómica estructural permite describir la cantidad y localización de genes, además, facilita la generación de modelos tridimensionales de macromoléculas codificadas en el genoma (S. E. Brenner, 2001). La genómica funcional ayuda a estudiar la función biológica de los genes, las proteínas y sus interacciones (López-López et al., 2005). La genómica comparativa se dedica a analizar características comunes y únicas entre genomas de diversas especies, incluso, es empleada en la anotación funcional de genomas recién secuenciados (Edwards & Holt, 2013; Koonin & Wolf, 2008; Sivashankari & Shanmughavel, 2007). También permite comparar la cantidad y localización de los genes (Liang, Zhao, Wei, Wen, & Qin, 2006); así como el contenido de GC (Hardison, 2003). Estudios importantes emplean a la genómica comparativa como perspectiva de estudio en bacterias para describir la conservación de la capacidad de patogenicidad (Wieland et al., 1994), de distribución filogenética de rutas metabólicas (Liang et al., 2006) y de evolución (Alföldi & Lindblad-Toh, 2013).

Dominios, motivos, familias y homología en proteínas

Las proteínas son biomoléculas formadas por cadenas lineales de aminoácidos que pueden adoptar una estructura tridimensional para llevar a cabo funciones como hormonas, transportadores, receptores, enzimas, entre otras (Alberts & Bray, 2006). En rutas metabólicas, como la biosíntesis de carotenoides, las proteínas enzimáticas permiten la aceleración de reacciones químicas debido a que disminuyen la energía de activación en las reacciones ($\Delta G<0$) (Lodish et al., 2000).

En la naturaleza, son 20 los aminoácidos principales que participan en la formación de proteínas. Cada aminoácidos tiene propiedades fisicoquímicas diferentes por lo que dependiendo de la secuencia de los aminoácidos será la función y estructura de las proteínas enzimáticas (Anfinsen, 1973; Armstrong, Alberti, Leach, & Hearst, 1989; Nelson & Cox, 2008a).

La estructura de las proteínas puede ser descrita a través de cuatro patrones estructurales: estructura primaria, secundaria, terciaria y cuaternaria (Nelson & Cox, 2008b; Richardson, 1981). La estructura primaria de una proteína describe la secuencia de aminoácidos (Nelson & Cox, 2008a); la estructura secundaria evidencia los patrones estructurales (*ej*. lámina beta y hélice alfa) formados por el arreglo de los aminoácidos y puentes de hidrógeno (Müller-Esterl, 2009; Nelson & Cox, 2008b; Voet & Voet, 2011a); la estructura terciaria consiste en la descripción tridimensional de una cadena polipeptídica (Anfinsen, 1973; Voet & Voet, 2011a); y la estructura cuaternaria consiste en el arreglo tridimensional de dos o más subunidades polipeptídicas (Nelson & Cox, 2008a).

Como se mencionó anteriormente, la secuencia de aminoácidos es un factor determinante en las estructuras y funciones de las proteínas, por lo que cuando se presentan proteínas con secuencias de aminoácidos equivalentes se pueden hacer aproximaciones sobre de sus propiedades. Dichos patrones conservados en la secuencia de aminoácidos son definidos como motivos y permiten aproximarse a la

función bioquímica y estructura de dicha proteína (Gupta & Lorenzini, 2007; Janin & Chothia, 1985; Rehm, 2001). Los motivos pueden ser descritos de diferente forma dependiendo del patrón estructural al que se refiera. Un ejemplo de un motivo en una estructura primaria es el motivo de dedo de zinc (CXX(XX)CXXXXXXXXXXXXHXXXH) (Bork & Koonin, 1996; Venugopal, Srinivasa, & Patnaik, 2009). En cambio, un motivo en una estructura secundaria es una región de una cadena polipeptídica que se relaciona con una función en particular o define una porción estructural de un dominio (Venugopal et al., 2009). De acuerdo con Nelson y Cox (2008), un motivo en una estructura secundaria también puede ser definido como una región de una cadena polipeptídica que permite el plegamiento de dos o más estructuras secundarias (*ej. beta-alfa-beta loop*). En ocasiones, pueden presentarse varios motivos consecutivos en una sola proteína. A estas regiones de varios motivos consecutivos se le conoce como *fingerprint* o bloque (Attwood & Beck, 1994).

En la estructura terciaria de las proteínas se encuentran regiones o subunidades modulares de aminoácidos que regulan su función y estructura (Janin & Chothia, 1985). Dichas regiones son llamadas dominios proteicos. Los dominios son definidos como una región de una cadena polipeptídica que es estable de forma independiente y determina una función o estructura particular en la proteína (Nelson & Cox, 2008b; Richardson, 1981).

La conservación de motivos estructurales y dominios permite agrupar a las proteínas en familias. Estas familias agrupan a proteínas que tienen similitud en su estructura primaria, estructura terciaria, función y conservan un ancestro común (Alberts & Bray, 2006; Bork & Koonin, 1996; Lodish et al., 2000). Cuando se presentan varias familias de proteínas con similitud en su secuencia de aminoácidos, así como también en sus principales motivos estructurales y similitudes funcionales, pueden ser agrupadas en superfamilias de proteínas (Nelson & Cox, 2008b). Pueden existir familias de proteínas con una amplia cobertura en los tres dominios (*i.e.* Bacteria, Archaea y Eucarya), lo cual

podría relacionarse con un origen común u homología (Daubin, Gouy, & Perrière, 2002a).

Las proteínas homólogas son aquellas que presentan similitud entre sus secuencias de aminoácidos y comparten un ancestro común (Voet & Voet, 2011b). Las proteínas homólogas suelen ser agrupadas en dos categorías: ortólogas y parálogas. Las proteínas ortólogas son proteínas homólogas que se localizan en diferentes especies y que se separaron por un evento de especiación (Brown, 2002); en cambio, las proteínas parálogas son proteínas homólogas que se presentan en el mismo organismo y que se originaron por un evento de duplicación de genes (Brown, 2002; Hardison, 2003). Las proteínas ortólogas pueden mantener la misma función en diferentes especies, en cambio, las proteínas parálogas pueden divergir gradualmente en su secuencia hasta diferenciarse en su función (Nelson & Cox, 2008b). Una categoría adicional de proteínas homólogas es conocida como xenología y se debe a la transferencia horizontal de genes (Klassen, 2010).

En proteínas de tipo enzimático, las proteínas pueden coincidir en su función debido a homología o por convergencia evolutiva. La convergencia consiste en caracteres que han evolucionado independientemente pero que coinciden en su estructura o función (Lodish et al., 2000). En el caso de las enzimas, la convergencia se presenta cuando dos enzimas tienen diferente secuencia de aminoácidos en su estructura primaria pero catalizan la misma reacción. A dichas enzimas se les conoce como isoenzimas (IUPAC-IUB Comission on Biochemical Nomenclature, 1974).

Bioinformática y búsqueda de homólogos

El término de bioinformática fue establecido por Paulien Hogeweg y Ben Hesper en 1978, como el estudio de los procesos informáticos en sistemas bióticos (Herbert et al., 2008; Hogeweg, 2011). Formalmente, la bioinformática es la disciplina encargada de analizar datos de origen biológico,

a través de herramientas asociadas a la informática (Luscombe, Greenbaum, & Gerstein, 2001; Rehm, 2001). Entre las principales aplicaciones que tiene la bioinformática destacan la organización de la información en bases de datos; el desarrollo de programas para el alineamiento múltiple de secuencias; la reconstrucción filogenética de las especies y la búsqueda de secuencias homólogas (Rehm, 2001).

Bases de datos

Las bases de datos bioinformáticas permiten recopilar, organizar y almacenar datos de origen biológico (*ej.* secuencias de nucleótidos o aminoácidos) (Rehm, 2001). Dichas bases de datos pueden ser primarias o secundarias, dependiendo de la fuente de la que se obtengan los datos (Herbert et al., 2008). Una base de datos primaria contiene secuencias de DNA, secuencias de proteínas y estructuras de proteínas obtenidas directamente de resultados experimentales, en tanto que, una base de datos secundaria es una base de datos curada, no redundante, que se obtiene a partir de bases de datos primarios (Herbert et al., 2008). Incluso, una base de datos secundaria puede contener datos sobre familias de proteínas, motivos o dominios proteínicos, familias de genes, mutaciones, etc. (Herbert et al., 2008).

Una de las principales bases de datos primaria es UniProtKB (*Universal Protein Resource Knowledgebase*) (www.uniprot.org). UniprotKB contiene información de proteínas integrando información principalmente de las bases de datos Swiss-Prot y TrEMBL (Apweiler et al., 2004; Herbert et al., 2008). Swiss-Prot es una base de datos con información extraída de la literatura y del análisis computacional realizado por curadores de bases de datos (Apweiler et al., 2004). TrEMBL almacena un conjunto de entradas que esperan por una anotación manual completa (Apweiler et al., 2004). El origen de las secuencias de proteínas en la base de datos UniprotKB es diversa y proviene de

los mayores repositorios de información biológica a nivel molecular: EMBL-Bank/GenBank/DDBJ, *Protein Data Bank* (PDB). Entre las principales bases de datos secundarias se encuentran las que permiten identificar motivos (PROSITE y eMOTIF); bloques o fingerprints (BLOCKS y PRINTS) y perfiles o modelos de Márkov (Profile y Pfam) (Attwood & Beck, 1994; Finn et al., 2014; Finn, Clements, & Eddy, 2011; Huang & Brutlag, 2001; Pietrokovski, Henikoff, & Henikoff, 1996; Sigrist et al., 2002).

Adicionalmente a las bases de datos de proteínas, existen bases de datos que permiten recopilar secuencias completas de genomas como *Genome* de NCBI (http://www.ncbi.nlm.nih.gov/genome). En dicho servidor se reúnen datos sobre la secuencia de genomas de los tres dominios (*ie.* Archaea, Bacteria, Eukarya), además de mapas genómicos y anotaciones funcionales (National Center for Biotechnology Information, 2014).

Alineamientos de secuencias

Dependiendo del número de secuencias comparadas se puede hablar de dos alineamientos distintos: los alineamientos pareados y los alineamientos múltiples de secuencias (AMS). Los alineamientos pareados consisten en sobreponer dos secuencias aparentemente similares y en evaluar la conservación entre sus nucleótidos (DNA, RNA) o aminoácidos (Baldi, Chauvint, Hunkapiller, & McClure, 1994; Lodish et al., 2000). La finalidad de los alineamientos por pares es inferir la homología entre secuencias similares y evidenciar patrones estructurales similares o funciones equivalentes (Altschul, 1998). La similitud entre las secuencias nos puede ayudar a inferir la homología entre secuencias debido a que es una variable cuantitativa; por esta razón se puede cuantificar la similitud pero no se puede cuantificar la homología de las secuencias. Es decir, las secuencias comparadas son

homólogas o no lo son.

El valor de similitud que existe entre las secuencias comparadas depende del *score* o calificación del alineamiento (Altschul, Gish, Miller, Myers, & Lipman, 1990). El *score* de alineamiento es la suma de los valores de sustitución para cada par de letras alineadas y *gaps* insertados (Altschul, 1998). Los *scores* de sustitución corresponden a valores asignados a cada par de letras alineadas y representan la probabilidad de sustitución de una base por otra (los valores de sustitución dependen de la matriz de sustitución empleada) (Henikoff & Henikoff, 1992).

Una matriz de sustitución reúne todos los posibles valores entre nucleótidos o aminoácidos alineados y describe el ritmo en el que un carácter de una secuencia cambia a otro carácter, con el tiempo (Henikoff & Henikoff, 1992). Las matrices más usadas en proteínas son las PAM (*Point Accepted Mutation*) y las BLOSUM (*BLOcks Substitution Matrix*) (Dayhoff & Schwartz, 1978; Henikoff & Henikoff, 1992; Yu & Altschul, 2005). PAM describe los cambios de un aminoácido a otro, que se han observado a lo largo de la evolución (*ej.* isoleucina por valina) (Dayhoff & Schwartz, 1978; Yu & Altschul, 2005). BLOSUM se basa en la conservación de bloques o dominios conservados (Henikoff & Henikoff, 1992). Los valores de BLOSUM se basan en la frecuencia de sustitución en bloques, de alineamientos locales. En matrices como BLOSUM, se pueden identificar a proteínas homólogas si presentan al menos 25-35% de identidad entre sus secuencias (Rehm, 2001). Los *gaps* (-) o huecos en el alineamiento representan eventos de mutación por inserciones o deleciones entre secuencias y permiten la correspondencia entre los sitios conservados (Sudha, 2014). Los *gaps* son penalizados con diferentes costos dependiendo de su longitud, entre más *gaps* tenga un alineamiento, la calificación de alineamiento será menor (A.Shehab, Keshk, & Mahgoub, 2012). Pueden existir diferentes alineamientos dependiendo del número de *gaps* en el alineamiento de secuencias. Para saber cuál es el alineamiento con mayor *score,* se debe de conocer la cantidad de residuos que coinciden, el

porcentaje de identidad *(ie.* número de coincidencias cada cien posiciones), el porcentaje de similitud *(ie.* similitud fisicoquímica de los aminoácidos), entre otras características relacionadas con la matriz de sustitución empleada (Altschul et al., 1990; Smith & Waterman, 1981; Yu & Altschul, 2005).

Existen dos tipos de métodos para generar alineamientos: alineamientos globales y alineamientos locales (Rehm, 2001). El alineamiento global consiste en alinear secuencias ocupando la longitud total de éstas, en tanto, el alineamiento local sólo alinea los fragmentos con mayor similitud (Altschul, 1998; Pietrokovski et al., 1996). Los alineamientos globales se basan en el algoritmo de Needleman-Wunsch; en cambio, los alineamientos locales se basan en el algoritmo de Smith-Waterman (Neddleman & Wunsch, 1970; Smith & Waterman, 1981). El algoritmo de Needleman-Wunsch busca la obtención del mejor alineamiento sin recurrir a todos los alineamientos disponibles entre dos secuencias, además, maximiza el número de coincidencias entre aminoácidos y minimiza la cantidad de *gaps* entre las secuencias (Neddleman & Wunsch, 1970). El algoritmo de Smith-Waterman permite localizar secuencias con mayor similitud en un región de su secuencia (Sudha, 2014). Un algoritmo de alineamiento local similar al de Smith-Waterman, es BLAST (*Basic Local Alignment Tool*). BLAST encuentra regiones similares entre pares de secuencias y calcula la significancia estadística de los aciertos (Altschul et al., 1990). Una de las versiones de BLAST aplicadas propiamente para proteínas es BLASTP debido a que permite comparar regiones de secuencias de aminoácidos e identificar secuencias similares en bases de datos (Kaur, Singh, & Singh, 2008).

Los alineamientos múltiples de secuencias (AMS) sirven para comparar múltiples secuencias y detectar regiones conservadas en secuencias de DNA o proteínas homólogas; tales regiones conservadas son asociadas a patrones estructurales y actividades bioquímicas características de una familia de proteínas (Eddy, 1998). Entre los programas más comunes para hacer AMS se encuentra MAFFT, que permite reducir el tiempo de cómputo de los AMS e incrementar su precisión al usar

métodos progresivos y después iterativos (Katoh, Misawa, Kuma, & Miyata, 2002; Katoh & Toh, 2008).

Búsqueda de homólogos por perfiles y Modelos Ocultos de Márkov

Los AMS también pueden ser empleados para buscar secuencias homólogas en bases de datos. A diferencia de la búsqueda a partir de una sola secuencia, los AMS permiten emplear modelos estadísticos (*ej.* secuencias consenso, patrones, matrices de puntuación de posiciones específicas, perfiles y Modelos Ocultos de Márkov) (Altschul, 1998; Frenz, 2008; Jones, 1999). Una secuencia consenso es el método más sencillo para construir un modelo estadístico, se basa en las frecuencias y representa el aminoácido más abundante en cada posición del alineamiento (Altschul, 1998).

Un patrón describe un conjunto de secuencias usando una simple expresión (expresión regular) (Frenz, 2008). El patrón usa el código de una letra para representar la posición de los aminoácidos, también usa los corchetes "[ ]" como ambigüedades, el símbolo "X" como representación de cualquier aminoácido y entre paréntesis "( )" se indica la repetición de caracteres (*ej.* .sitio activo de L-lactato deshidrogenasa <[LIVMA]-G-[EQ]-H-G-[DN]-[ST]> (Frenz, 2008). Las desventajas de usar patrones radican en que es un modelo pobre para representar inserciones o deleciones, además, pequeños patrones son más proclives a encontrar falsos positivos.

Una matriz de puntuación de posiciones específicas o *position-specific scoring matrix* (PSSMs) permite representar la frecuencia de los residuos de forma probabilística en cada columna del alineamiento (Jones, 1999). Las PSSMs pueden ser usadas como una secuencia para realizar búsquedas y alineamientos en bases de datos. Una desventaja de los PSSMS deriva cuando se presentan *gaps* en el alineamiento múltiple de secuencias, ya que los PSSMs suelen no funcionar en estas situaciones debido

a que no contemplan los huecos (Jones, 1999). Debido a lo anterior, se prefieren usar perfiles de proteínas.

Un perfil es similar a una PSSM sólo que en el perfil se agrega información sobre penalización por *gaps* y se pueden usar para detectar secuencias homólogas remotas (secuencias que han divergido a lo largo del tiempo y que por lo tanto presentan bajos niveles de similitud, en un rango de identidad entre secuencias de 20 a 30%. Dichas secuencias homólogas remotas, aún conservan los principales patrones estructurales y funcionales que corresponden a una familia de proteínas) (Logan, Moreno, Suzek, Weng, & Kasif, 2001; Pearson, 2013; Rost, 1999). PSI-BLAST (*Position-specific iterated* BLAST) permite la construcción de perfiles para después realizar búsquedas de forma iterativa en una base de datos (Altschul et al., 1997). La ventaja de usar perfiles de PSI-BLAST es que permite encontrar más secuencias homólogas que un clásico BLAST, sin embargo, los perfiles de PSI-BLAST pueden ser poco selectivos después de varias iteraciones pudiendo encontrar falsos positivos en la construcción gradual del perfil (Altschul et al., 1997). Como alternativa a los perfiles de PSI-BLAST se sugiere usar perfiles de Modelos Ocultos de Márkov o *Hidden Markov Models* (HMM) debido a que mejoran la calidad del modelo aún después de varias iteraciones. Para poder entender los perfiles HMM, se necesitan describir inicialmente las Cadenas de Márkov.

Cadenas de Márkov

Una Cadena de Márkov es una sucesión de variables aleatorias en donde la probabilidad de que ocurra una variable depende de su variable predecesora (propiedad de Márkov) (Prada-Alonso, 2013). Se le conoce como "cadena" debido a que es una serie de eventos consecutivos que se encuentran enlazados por probabilidades de transición (Prada-Alonso, 2013). La propiedad markoviana determina

que un evento en el futuro dependa únicamente de lo que ocurra en el presente, sin importar el pasado (Baldi et al., 1994; Prada-Alonso, 2013).

Las Cadenas de Márkov están compuestas por variables o estados (M), y probabilidades de transición (P). Las variables representan elementos particulares en un momento indicado y la probabilidad de transición describe la probabilidad de que un estado cambie al estado siguiente (Prada-Alonso, 2013). En la **Figura 1, A**, se representa una Cadena de Márkov simplificada en la que se observan dos variables (M1 y M2) entrelazadas por probabilidades de transición (P). El sentido del modelo transcurre en un tiempo discreto pasando por cada variable hasta llegar a la variable "final" del modelo. El sistema comienza cuando la variable inicial es definida con algún elemento cualquiera, lo que favorece que se emita una probabilidad de transición equivalente al elemento definido. Continuo a esto, la probabilidad de transición permite que la primer variable sea conocida y sea definida como M1. Cuando la variable M1 es conocida se emite una probabilidad de transición para que el siguiente estado sea definido. En este caso la nueva variable es mostrada y es definida como M2. Cuando la variable M2 es definida, se emite una probabilidad de transición a un estado final. En este momento, la propiedad markoviana se hace presente debido a que la variable "final" dependió unicamente de su estado inmediato anterior (M2) y no el resto de las variables que les antecedieron (inicio, M1). Cabe mencionar que los estados M1 y M2 son visibles debido a que se conocen los elementos que representan. En este caso, las variables M pueden ser representaciones de nucleótidos o aminoácidos.

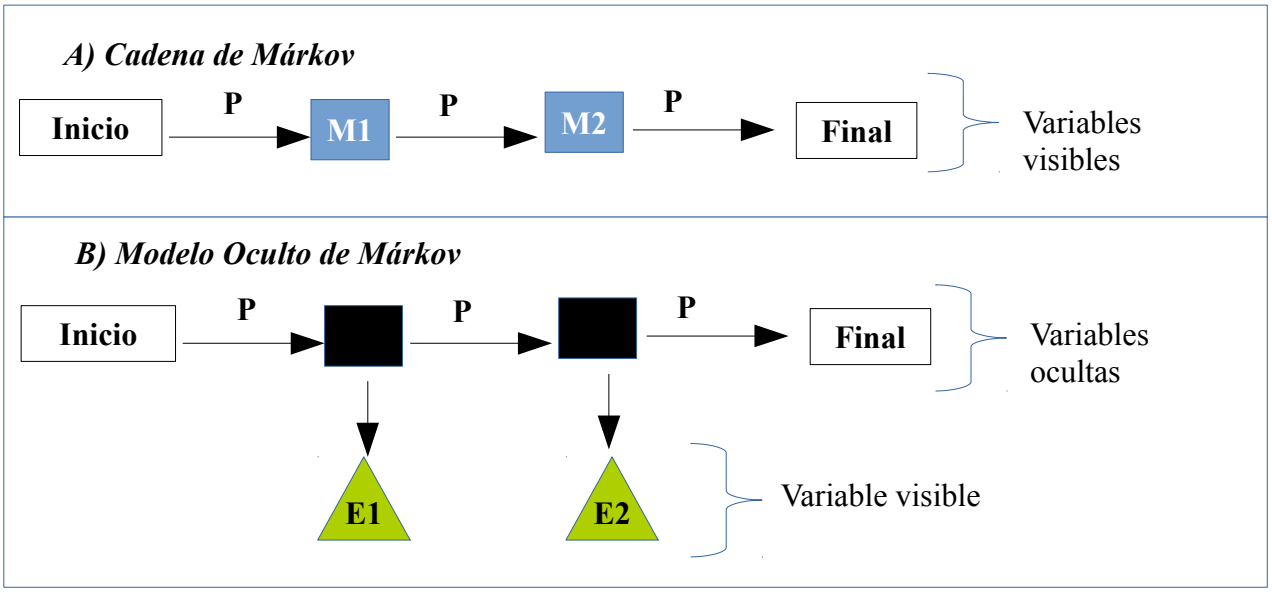


Figura 1. Diagrama simplificado de una Cadena de Márkov (A) y un Modelo Oculto de Márkov (B). Las variables o estados son representados en cuadrados (M1 y M2) y las probabilidades de transición se ejemplifican con la letra "P". Para el caso particular de los Modelos Ocultos de Márkov se muestran las variables de emisión como triángulos (E1 y E2). La presencia de variables ocultas es la principal diferencia entre un Modelo Oculto de Márkov y una Cadena de Márkov

Modelos Ocultos de Márkov

Los Modelos Ocultos de Márkov o *Hidden Markov Models* (HMM) son modelos probabilísticos que permiten la modelación de familias de proteínas, la predicción o búsqueda de dominios y la búsqueda de homólogos (Dai & Cheng, 2008). Los HMM fueron descritos por Leonard Baum, a finales de los 60's y debido a su capacidad para predecir eventos o variables ocultas, los HMM son usados

actualmente para el reconocimiento de patrones de habla, criptoanálisis y bioinformática (Baum & Eagon, 1967).

En bioinformática, los HMM generan un sistema de puntaje de posiciones específicas basado en la frecuencia de los residuos de un alineamiento múltiple de secuencias y en matrices de sustitución (*ej.* BLOSUM62) (Eddy, 2008). Los perfiles HMM facilitan la búsqueda de homólogos remotos ya que no dependen de un algoritmo de alineamiento de secuencias, sino que su eficiencia depende de las secuencias con las que se entrene el perfil y a diferencia de BLAST, los perfiles HMM son más eficientes en el reconocimiento de homólogos remotos ya que usan múltiples secuencias alineadas (Eddy, 1998; Nahas, Kassim, & Shikoun, 2012). A diferencia de una PSSM, los HMMs dependen de matrices de sustitución basadas en modelos evolutivos (*ej.* BLOSUM62) (Bykova, Favorov, & Mironov, 2013). Y en comparación con PSI-BLAST, los perfiles HMM son eficientes desde su primer entrenamiento y suelen ser más selectivos que los perfiles generados por PSI-BLAST (Hoberman & Durand, 2011; Mona & Parker, 1999).

A diferencia de las cadenas de Márkov, en los HMMs se cuenta con variables ocultas (**Figura 1**) (Baum & Eagon, 1967). Las variables ocultas consisten en eventos que cumplen la propiedad de Márkov pero que no son observables o conocidas. Cada variable oculta representa una distribución de probabilidad que es independiente entre variables. En la distribución de probabilidad se aloja la probabilidad de cada uno de los 20 aminoácidos posibles que forman a las proteínas. Entre mayor frecuencia tenga un aminoácido en un AMS, mayor será su probabilidad en los estados ocultos (Karplus, Barrett, & Hughey, 1998; Mona & Parker, 1999). En caso de que un aminoácido no esté presente en el AMS, el HMM le asigna la probabilidad de 0.01 (1%) como valor predeterminado. La probabilidad de un aminoácido no puede ser incrementada sin reducir la probabilidad de uno o más aminoácidos (Eddy, 1998). Cabe mencionar que la probabilidad de un aminoácido no sólo depende de

la abundancia con la que se presente en dicha posición, sino que también depende de su probabilidad de transición a otro aminoácido (basada en matrices generales como BLOSUM62) (Finn et al., 2011; Henikoff & Henikoff, 1992; Pietrokovski et al., 1996). La matriz de sustitución BLOSUM62 es empleada en HMM, debido a su capacidad para definir la probabilidad de sustitución entre aminoácidos basado en lo observado durante la evolución (Henikoff & Henikoff, 1992).

Los perfiles HMM son HMM aplicados a la representación de familias de proteínas y en estos no sólo existen transiciones entre estados de aminoácidos, sino que también existen probabilidades de transición para estados de inserción o deleción (**Figura 2**) (Eddy, 1998; Hoberman & Durand, 2011).

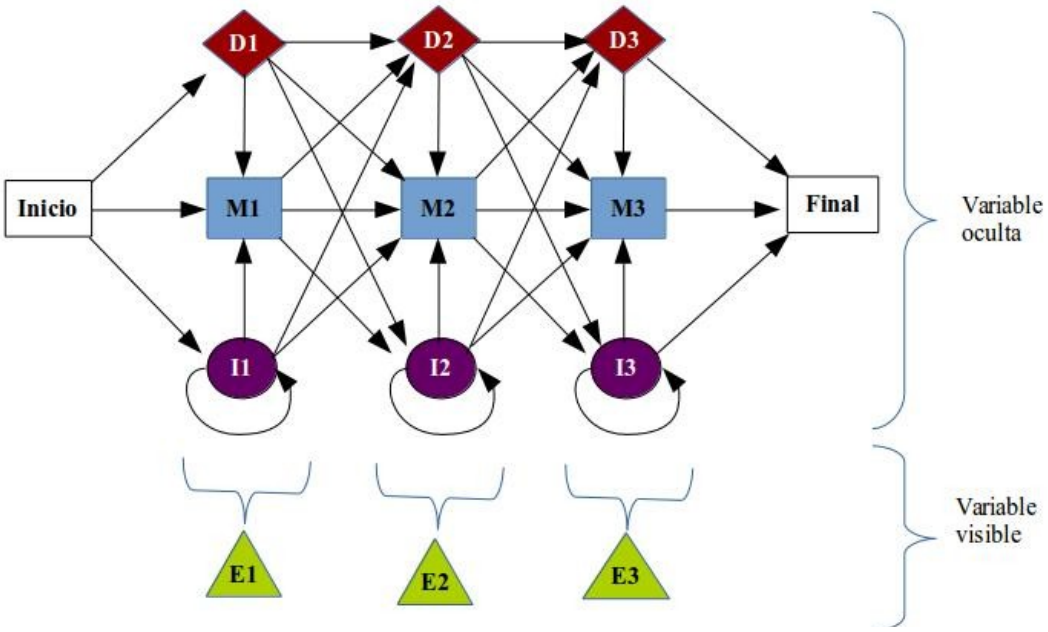


Figura 2. Representación de un perfil de Modelo Oculto de Márkov o *Hidden Markov Model* (HMM) con estados de inserción y deleción. En esta figura, la variable oculta es representada con diferentes variables (*ie*., estados de deleción en rombos D1, D2, D3; estados con distribución de probabilidad en cuadrados M1, M2, M3; estados de inserción en círculos I1, I2, I3). La variable visible es representada por los estados de emisión en triángulos (E1, E2, E3).

En el entrenamiento de los perfiles HMM, la probabilidad de una inserción o deleción dependen de la proporción de *gaps* en el alineamiento (Finn et al., 2011; Hoberman & Durand, 2011). La probabilidad de transición a un estado de inserción depende de la frecuencia de *gaps* que se presente en dicha posición (**Figura 3**). La transición a un estado M, de inserción o de deleción, generan probabilidades de emisión distintas que permiten cuantificar la calidad del alineamiento entre secuencias y el perfil HMM (Hoberman & Durand, 2011).

Cuando se compara una secuencia con un perfil HMM, el perfil HMM elige la mejor ruta para cubrir cada aminoácido de la secuencia y pasa de estado a estado generando una probabilidad de emisión por cada paso (**Figura 4**) (Eddy, 1998; Finn et al., 2011). Entre más coincida la secuencia con el perfil HMM, mayor probabilidad de emisión tendrá cada estado visible. La suma de probabilidades de emisión permite encontrar la mejor ruta para dicha secuencia. Para conocer cuantitativamente el mejor alineamiento, los perfiles HMM son capaces de reportar un *E-value* y un *bit score*.

De acuerdo con Finn (*et al.,* 2011), un *bit score* de perfiles HMM, es un puntaje proporcional *log-odds* (base dos) que compara la probabilidad del perfil HMM contra la probabilidad de una hipótesis nula (un modelo de secuencia aleatoriamente distribuido e independiente, como en BLAST) (Eddy, 2008). El *E-value* corresponde a los aciertos esperados al azar cuando se busca una secuencia en una base de datos de secuencias no homólogas (Eddy, 2008). Entre más grande sea el *E-value*, menor calidad tendrá el alineamiento. Ya que el *E-value* depende del tamaño de la base de datos en la que se busca una secuencia y de la longitud de las secuencias que se estén analizando, el *E-value* no siempre es un valor usado al comparar alineamientos y alternativamente se usa el *bit score* ya que este no depende del tamaño de la base de datos comparada (Eddy, 2008; Finn et al., 2011). Entre más grande sea el *bit score,* mejor será el alineamiento entre el perfil HMM y la secuencia (Bykova et al., 2013; Eddy, 1998).

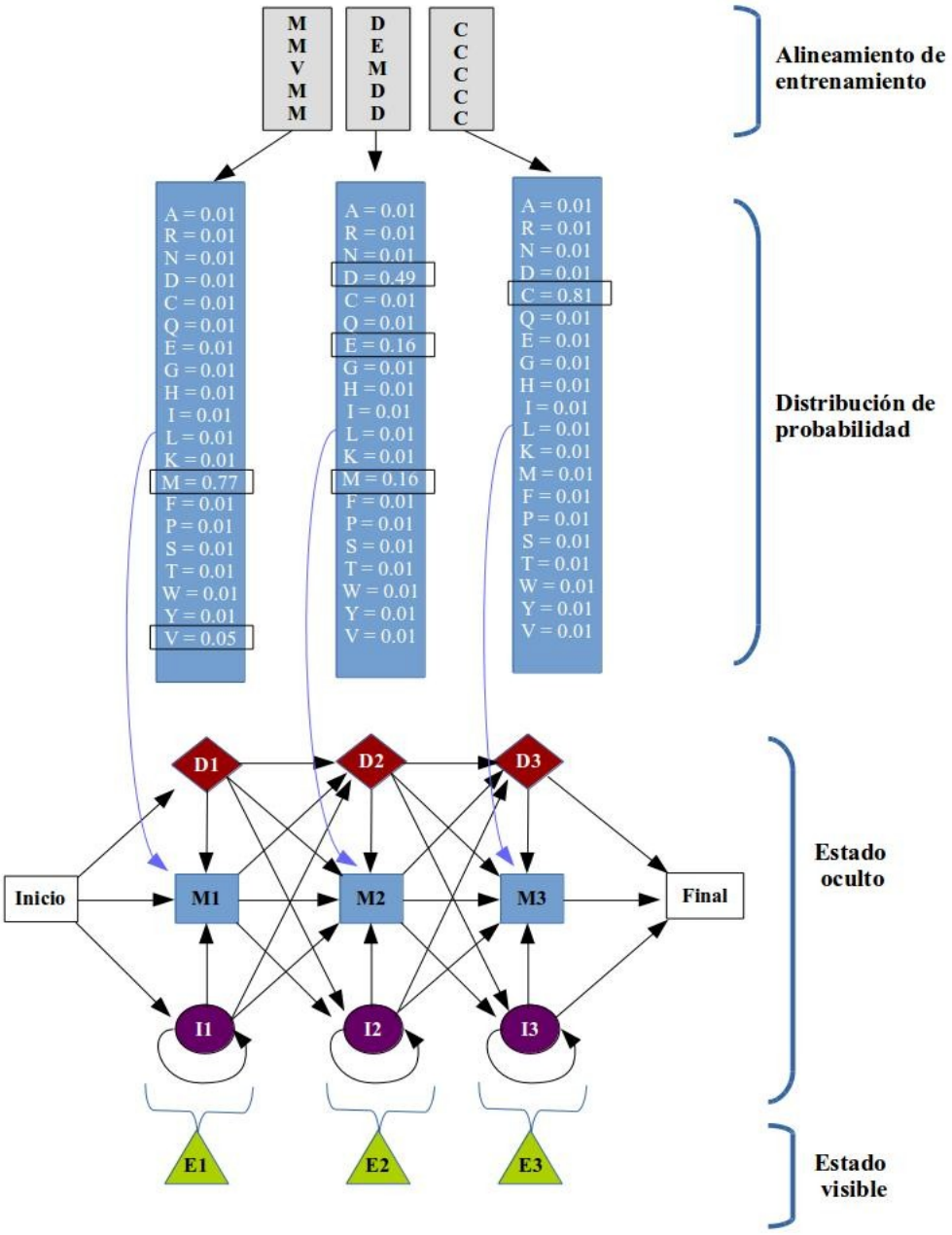


Figura 3. Entrenamiento de un perfil HMM a partir de un alineamiento múltiple de secuencias. En la parte superior se encuentra dicho alineamiento. En la parte media se observan tres columnas que contienen la distribución de probabilidad de los 20 aminoácidos posibles. En la parte inferior se muestra el perfil HMM con sus variables ocultas y visibles.

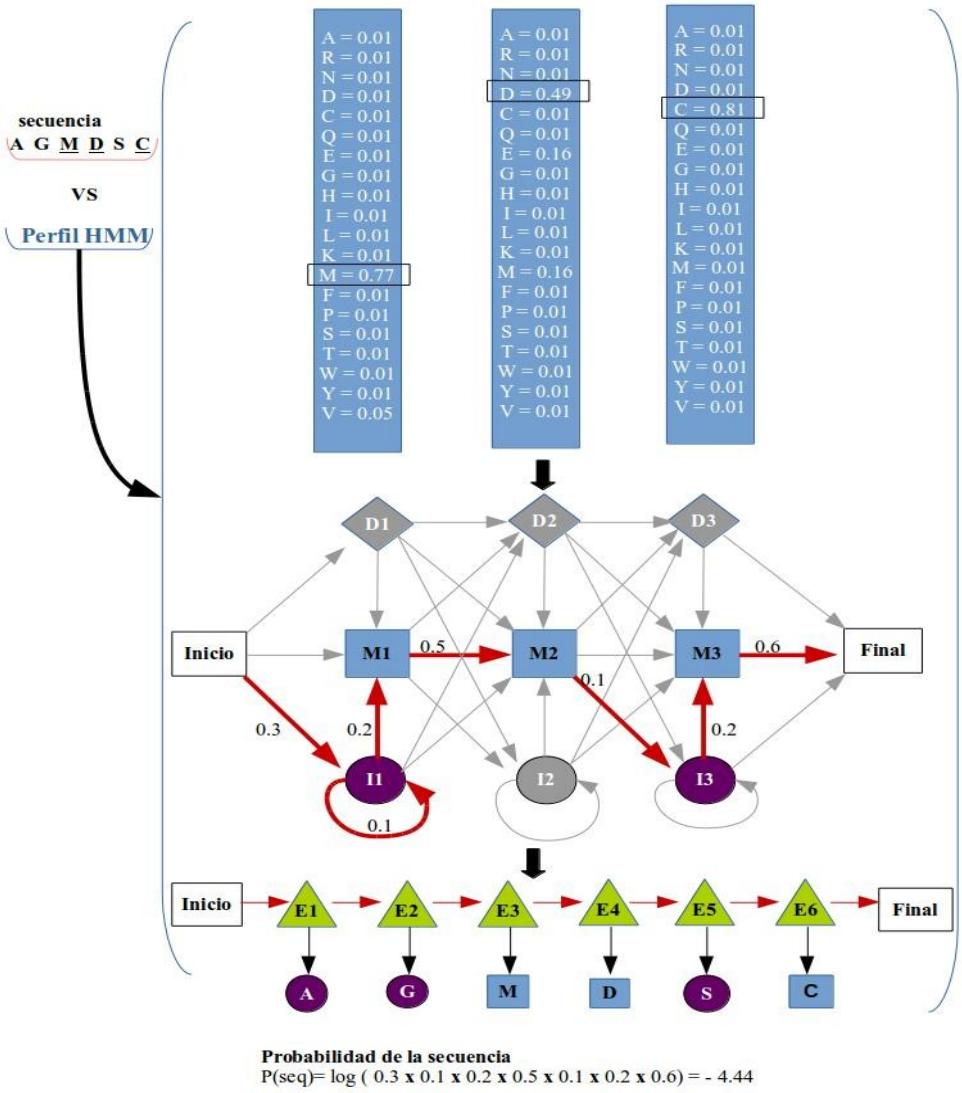


Figura 4. Alineamiento entre una secuencia y un perfil HMM. El perfil HMM evalúa sistemáticamente cada posición de la secuencia y las clasifica en cualquiera de los tres estados *(ie.* inserción (I), deleción (D) o coincidencia (M)). Por cada transición, el modelo genera una variable oculta y una variable visible. La variable oculta influye en la transición al siguiente estado y la variable visible se muestra como las probabilidades de emisión. Entre mayor sea la similitud entre una secuencia y el modelo, mayor será la probabilidad de emisión.

En el caso de la **Figura 4**, el perfil HMM evalúa cada posición de la secuencia y le asigna sólo un estado (M, I o D), dependiendo de su similitud. Si el primer aminoácido de la secuencia (A= alanina) no coincide con el aminoácido de mayor probabilidad (M= metionina), del estado inicial del modelo (M1), entonces se emite un estado de inserción y una probabilidad de emisión equivalente. A partir del estado de inserción emitido, el modelo parte para conocer el siguiente estado. Si el segundo aminoácido de la secuencia tampoco coincide con el aminoácido de mayor probabilidad del estado, se emite otro estado de inserción hasta que se encuentre un aminoácido que coincida con el aminoácido modelo. Cuando el modelo encuentra una posición que coincida con la secuencia, entonces se genera una probabilidad de tipo M y así sucesivamente.

Existen diversos programas para el entrenamiento de los perfiles como SAM, HMMpro y HMMer (Finn et al., 2011). También existen programas para la visualización de perfiles de Márkov mediante un HMM logo, como el programa Skylign (Wheeler, Clements, & Finn, 2014). Un HMM logo permite visualizar la información contenida y la distribución de proteínas de un alineamiento múltiple de secuencias (T. D. Schneider & Stephens, 1990). En los HMM logos se pueden observar columnas con letras apiladas de diferentes alturas, las letras representan los diferentes aminoácidos. Una altura superior representa mayor información del aminoácido en dicha posición, en la familia de proteínas (**Figura 5**) (Schuster-böckler, Schultz, & Rahmann, 2004). La altura de las columnas muestra que tan significativa es la probabilidad de emisión de un estado, respecto a la distribución de probabilidades dadas por el perfil completo *(ie.* entropía o información contenida) (T. D. Schneider & Stephens, 1990).

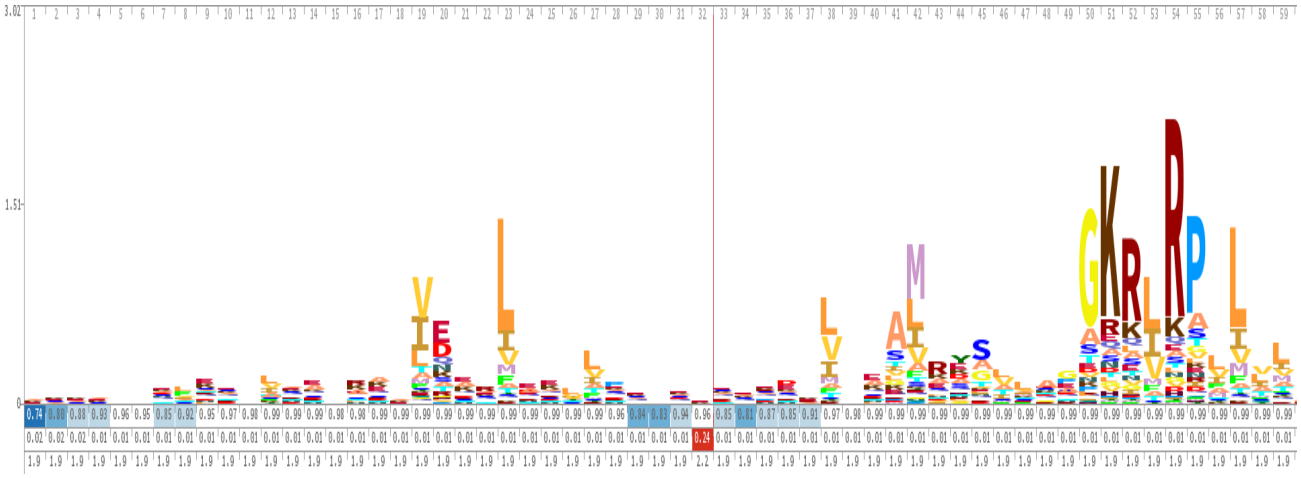


Figura 5. Ejemplo de un HMM logo de la proteína CrtE, posición 1-58. Los aminoácidos más conservados en el alineamiento múltiple de secuencias son mostrados en mayor tamaño y representan algunos posibles motivos conservados en la familia de proteínas. La escala vertical cuantifica la entropía o información contenida en cada posición. La tabulación inferior muestra las probabilidades de transición entre las posiciones. Los espacios sin letras representan estados de inserción.

La filogenómica y el árbol de la vida de Ciccarelli

La filogenómica permite inferir relaciones filogenéticas y estudiar los mecanismos de evolución de los organismos a partir de datos genómicos (Eisen & Fraser, 2003). A diferencia de la filogenética molecular, la filogenómica compara cientos de genes con la finalidad de describir con mayor resolución las relaciones filogenéticas entre las especies (Eisen & Fraser, 2003). Durante muchos años, la filogenética molecular ha sido la base para la generación de hipótesis sobre la historia evolutiva de diversos organismos. Uno de los casos más sobresalientes fue realizado por Woese y colaboradores en 1970, en el cual generaron un árbol filogenético universal basado en las secuencias de subunidades ribosomales pequeñas y lograron describir tres dominios celulares (*ie*. Bacteria, Eukarya y Archaea)

(Woese & Fox, 1977). Sin embargo, en ocasiones la historia de un solo gen no describe en su totalidad la historia evolutiva de las especies. En este sentido, diversos estudios han optado por usar múltiples genes homólogos para realizar hipótesis más detalladas que permitan la reclasificación taxonómica de taxones particulares (Ciccarelli et al., 2006; Daubin, Gouy, & Perrière, 2002b; Sentausa & Fournier, 2013).

En la actualidad se disponen más de 35 000 proyectos de secuenciación de genomas en bacterias, lo cual permite conocer no sólo la historia evolutiva de un solo gen, sino la historia evolutiva del conjunto de genes homólogos conservados en muchas especies de los tres dominios celulares (*ie*., Eukarya, Bacteria y Archaea). Un ejemplo claro de la aplicación de la filogenómica es el árbol filogenético propuesto por Ciccarelli *et al*., (2006). Ciccarelli generó una filogenia universal basada en la concatenación de 31 familias de genes ortólogos que se presentan en 191 especies (con genomas totalmente secuenciados), representantes de los tres dominios. Cabe mencionar que, Ciccarelli omitió los genes que hubieran padecido eventos de transferencia horizontal de genes y aquellos que presentaran complicaciones para ser alineados para brindarle mayor resolución a la topología de la filogenia. En su publicación, Ciccarelli destaca la importancia de usar múltiples marcadores evolutivos en genomas secuenciados para resolver las incertidumbres evolutivas entre especies de bacterias y la clasificación taxonómica de diversos grupos (Ciccarelli et al., 2006).

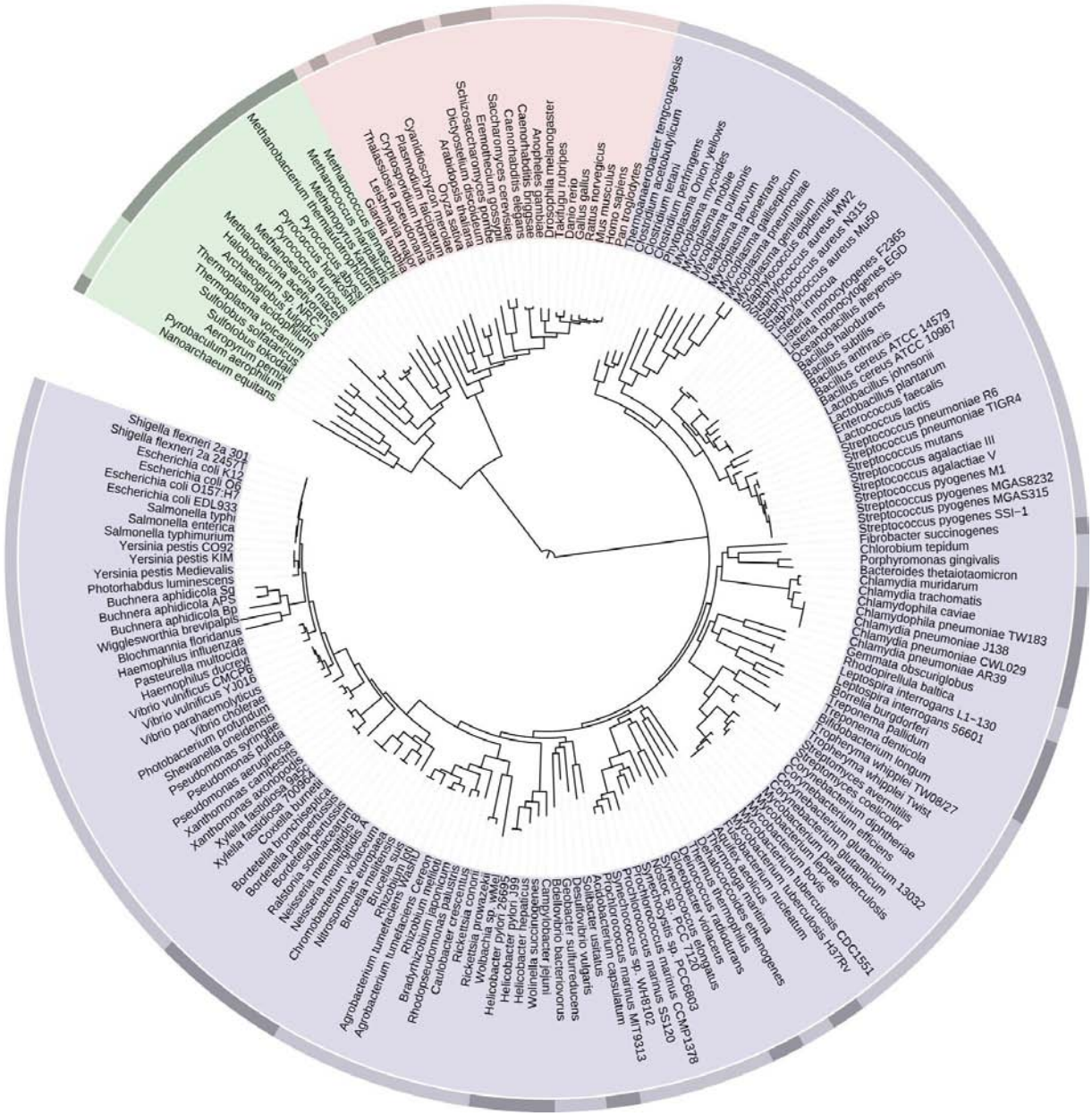


Figura 6. Filogenia propuesta por Ciccarelli (*et al*., 2006). En verde se muestran a los eucariontes, en rosa a arqueobacterias y en morado a eubacterias Disponible en http://itol.embl.de/itol.cgi (Letunic & Bork, 2011).

Los carotenoides y su importancia

Los carotenoides son metabolitos secundarios producidos por hongos, plantas, algas y bacterias (Arrach, Schmidhauser, & Avalos, 2002; F. X. Cunningham & Gantt, 1998; Takaichi, 2011). Dichos metabolitos, son compuestos derivados de los lípidos, que al ser hidrofóbicos pueden encontrarse asociados a membranas o a pigmentos como la bacterioclorofila o la clorofila, en organismos fotosintéticos (Armstrong, 1997; Meléndez-Martínez, Vicario, & Heredia, 2007).

En bacterias, los carotenoides son importantes por la diversidad de actividades que realizan. En bacterias fotosintéticas como lo son las cianobacterias, los carotenoides se relacionan principalmente con la captación de luz, la disipación del exceso de energía capturada, la transferencia de energía dentro del aparato fotosintético, la prevención del daño fotooxidativo y la regulación de la fluidez en la membrana (Francis X Cunningham, Sun, Chamovitz, Hirschberg, & Gantt, 1994; Liang et al., 2006; Meléndez-Martínez et al., 2007; Umeno, Tobias, & Frances, 2005). En cambio, en bacterias no fotosintéticas como *Bacillus safensis*, los carotenoides se relacionan particularmente con la esporulación, resistencia a la radiación UV y protección del DNA ante las especies reactivas de oxígeno (Khaneja et al., 2010a; Moeller, Horneck, Facius, & Stackebrandt, 2005; Perez-Fons et al., 2011).

La estructura química de los carotenoides es reconocida por presentar una región de polieno con siete enlaces dobles conjugados, denominada cromóforo. La región del cromóforo, permite a los carotenoides absorber longitudes de onda entre 400 y 700 nm y emitir coloraciones entre amarillo y rojo (Armstrong, 1997; Rodríguez Villalón, 2010; Takano, Obitsu, Beppu, & Ueda, 2005). En caso de que los carotenoides se asocien con proteínas, el rango de absorción puede extenderse a colores azul, púrpura y verde (Meléndez-Martínez et al., 2007). Para que los carotenoides presenten una coloración en el rango de radiación visible se necesitan, al menos, siete dobles enlaces en la región del polieno

(Meléndez-Martínez et al., 2007). La región del polieno, además de favorecer la absorción y emisión de radiación, también permite que los carotenoides tengan propiedades antioxidantes (Jáuregui-Carranco, Calvo-Carrillo, & Pérez-Gil-Romo, 2011; Paiva & Russell, 1999). La propiedad antioxidante de los carotenoides se debe a que la región de los dobles enlaces disminuye el estado de excitación de las moléculas radicales. Algunos de los radicales que los carotenoides pueden inactivar son el oxígeno (O_2) o -radicales peroxilo (ROO*) (Cogdell et al., 2000; Paiva & Russell, 1999).

La estructura química de los carotenoides está compuesta por ocho unidades de isoprenos, que en conjunto, forman una cadena de 40 átomos de carbono (Umeno et al., 2005). No obstante, también existen carotenoides que presentan estructuras químicas con 30 átomos de carbono; como el glicosil apolicopeno que es producido por especies de *Bacillus* formadoras de esporas (*ej. B. indicus*) (Moeller et al., 2005; Perez-Fons et al., 2011). Estos carotenoides con menos de 40 átomos de carbono son conocidos como apocarotenoides. En este sentido, también existen carotenoides con 40 o más átomos de carbono que son conocidos como homocarotenoides (Umeno, Tobias, & Arnold, 2002). Entre los homocarotenoides más conocidos, destaca la decaprenoxantina (C_{50}) producida por *Corynebacterium glutamicum* (Heider, Peters-Wendisch, & Wendisch, 2012). Los apocarotenoides son formados por los precursores de los homocarotenoides. En tanto que, los homocarotenoides pueden ser formados a partir de carotenoides con 40 átomos de carbono como el licopeno (Heider et al., 2012; Umeno et al., 2002).

Los carotenoides también pueden ser clasificados como xantófilas o carotenos, dependiendo de la presencia de un anillo ciclohexano oxigenado insaturado. En carotenoides tipo xantófila, el ciclohexano se encuentra oxigenado. En cambio, en los carotenoides tipo caroteno, el ciclohexano no se encuentra oxigenado (Jáuregui-Carranco et al., 2011). La luteína y el betacaroteno, son algunos de los carotenoides tipo xantófila y caroteno, respectivamente (von Lintig, 2010) (**Figura 7**).

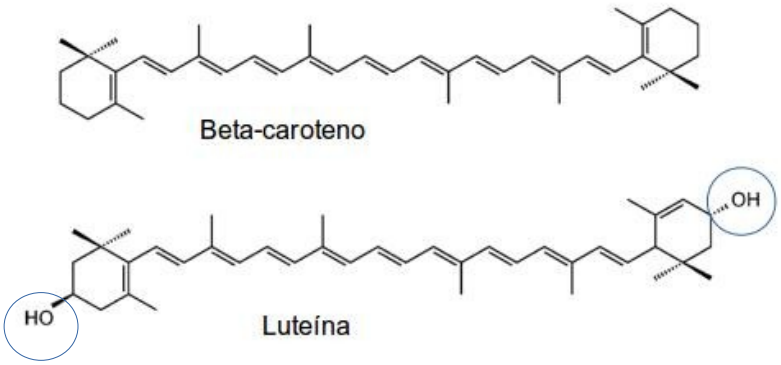


Figura 7. Estructura química del caroteno beta-caroteno y de la xantófila luteína. El grupo hidroxilo que caracteriza a las xantófilas se muestra encerrado en un círculo. Imagen modificada de von Lintig (2010).

Ruta de biosíntesis de carotenoides (BC)

La biosíntesis de carotenoides suele ser clasificada en las siguientes reacciones principales: biosíntesis del isopreno, síntesis de diapofitoeno, síntesis de fitoeno, deshidrogenación del fitoeno, ciclización del licopeno, reacciones posteriores a la ciclización del licopeno y formación de carotenoides acíclicos (Klassen, 2010; Phadwal, 2005; Sieiro, Poza, de Miguel, & Villa, 2003) (**Figura 8**).

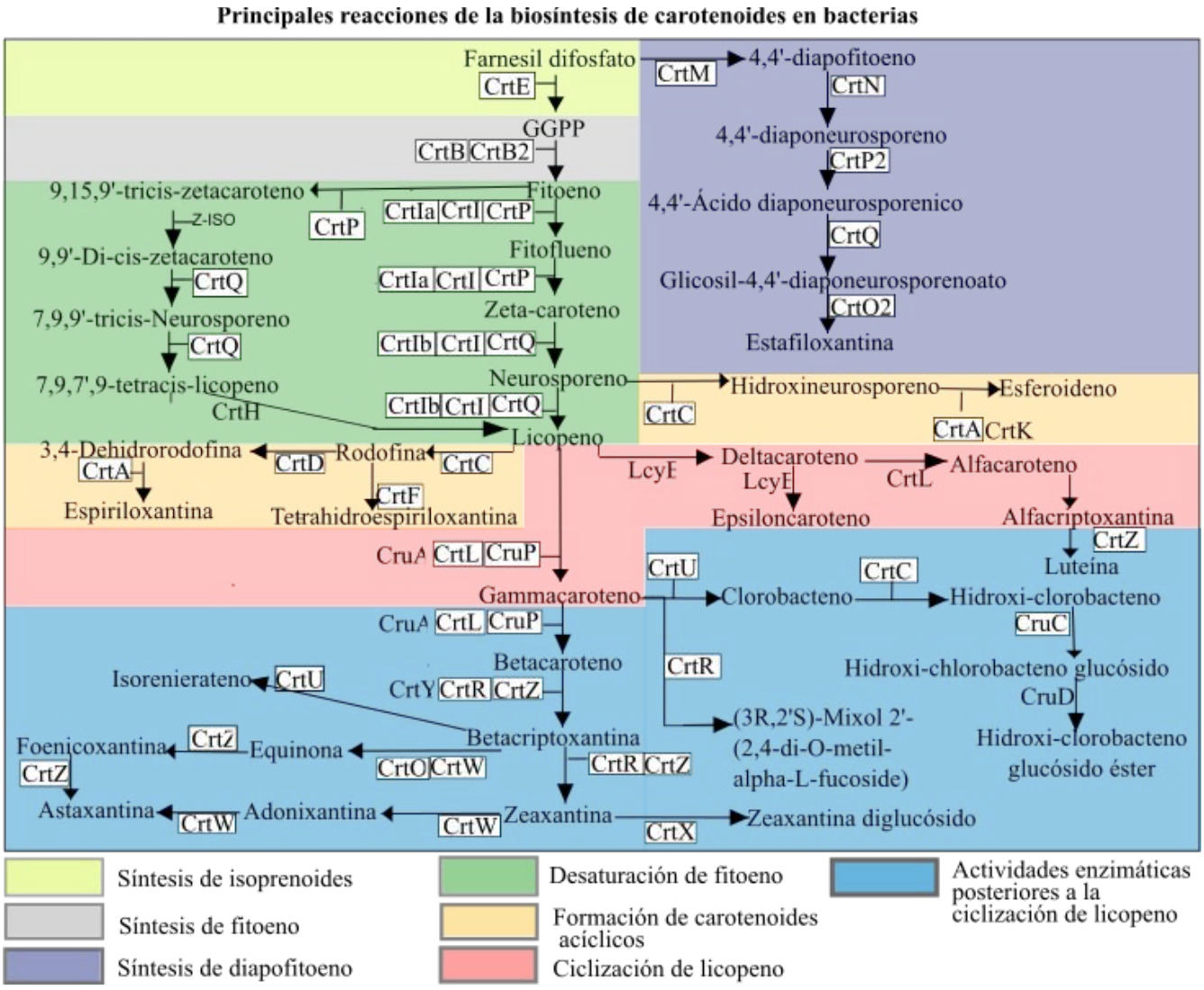


Figura 8. Ruta simplificada de la biosíntesis de carotenoides en bacterias. Se muestran las principales reacciones y en recuadros blancos se muestran las proteínas que actúan en cada reacción. Las proteínas mostradas, se describen en la **Tabla 1**.

Tabla 1. Principales proteínas de la biosíntesis de carotenoides. Las columnas contienen información sobre el proceso en el que están implicadas las proteínas, su la abreviación, la función que realizan las proteínas; la secuencia de referencia que fue usada para buscar más homólogos y las referencias bibliográficas. Las secuencias de referencia con asterisco (*) fueron obtenidas de UniProtKB/Swiss-Prot, las secuencias sin asterisco fueron obtenidas de UniProtKB/TrEMBL.

Reacciones principales de la BC	Proteína	Nombre de la enzima	Secuencia de referencia (UniProtKB/ SwissProt*)	Referencia
Biosíntesis de isoprenos	CrtE	Geranilgeranil difosfato sintasa [EC=2.5.1.29]	P22873*	(Shivanand, Hearst, & Poulter, 1992)
Síntesis de diapofitoeno	CrtM	Dehidroscualeno sintasa [EC=2.5.1.96]	Q2FV59 *	(Wieland et al., 1994)
Síntesis de diapofitoeno	CrtN	Dehidroscualeno desaturasa [EC=1.3.8.2]	O07855 *	(Wieland et al., 1994)
Síntesis de diapofitoeno	CrtP2	Diapolicopeno oxigenasa [EC=1.14.99.44]	Q2FV57*	(Tao et al. 2005)
Síntesis de diapofitoeno	CrtQ2	4'4-diaponeurosporenoato glicosiltransferasa [EC=2.4.1.-]	Q53590*	(Pelz et al., 2005)
Síntesis de diapofitoeno	CrtO2	Glicosil-4,4'-diaponeurosporeonato aciltransferasa [EC=2.3.1.-]	Q2YWE4*	(Pelz et al., 2005)
Síntesis de fitoeno	CrtB	Fitoeno sintasa [EC:2.5.1.32]	P54905 *	(Hoshino, Fujii, & Nakahara, 1993)
Síntesis de fitoeno	CrtB2	Proteína bifuncional (fitoeno sintasa/isoprenil transferasa) [EC:2.5.1.32] / [EC= 2.5.1.-]	Q9ACU1 *	(Heider et al., 2012)

Reacciones principales de la BC	Proteína	Nombre de la enzima	Secuencia de referencia (UniProtKB/ SwissProt*)	Referencia
Deshidrogenación del fitoeno	CrtP	15-cis-fitoeno desaturasa [EC=1.3.5.5]	P26294*	(Lopes et al., 2009)
Deshidrogenación del fitoeno	CrtQ	Zeta-charoteno desaturasa [EC=1.3.5.6]	P74306*	(C. Schneider, Böger, & Sandmann, 1997)
Deshidrogenación del fitoeno	CrtI	Fitoeno desaturasa formadora de neurosporeno [EC= 1.3.99.28]	P54980 *	(Iniesta, Cervantes, & Murillo, 2007)
Deshidrogenación del fitoeno	CrtIa	Fitoeno desaturasa formadora de zeta-caroteno [EC=1.3.99.29]	P54979*	(Iniesta, Cervantes, & Murillo, 2007)
Deshidrogenación del fitoeno	CrtIb	All-trans-zeta-caroteno desaturasa [EC=1.3.99.26]	Q02861*	(Iniesta et al., 2007)
Ciclización de licopeno	CrtL	Licopeno beta-ciclasa [EC:5.5.1.19]	Q55276*	(Iniesta et al., 2007)
Ciclización de licopeno	CruP	Licopeno ciclasa CruP [EC: 5.5.1.19]	A5A546	(Francis X Cunningham et al., 1994)
Reacciones posteriores a la ciclización de licopeno	CrtR	Beta-caroteno hidroxilasa [EC:1.14.13.-]	Q7V4X0	(Maresca, Graham, Wu, Eisen, & Bryant, 2007)
Reacciones posteriores a la ciclización de licopeno	CrtZ	Beta-caroteno 3-hidroxilasa [EC:1.14.13.129]	Q9LTG0*	(Choi, Matsuda, Hoshino, Peng, & Misawa, 2006)
Reacciones posteriores a la ciclización de	CrtU	Isorenierateno sintasa [EC:1.-.-.-]	V7KBF6	(Choi, Matsuda, Hoshino, Peng, & Misawa, 2006)

Reacciones principales de la BC	Proteína	Nombre de la enzima	Secuencia de referencia (UniProtKB/ SwissProt*)	Referencia
licopeno				
Reacciones posteriores a la ciclización de licopeno	CrtW	Beta-caroteno ketolasa, tipo CrtW [EC=1.13.-.-]	P54972*	(Tsuchiya et al., 2005)
Reacciones posteriores a la ciclización de licopeno	CrtX	Zeaxanthin glucosiltransferasa [EC: 2.4.1.276]	Q01330*	(Misawa et al., 1990)
Reacciones posteriores a la ciclización de licopeno	CruC	Clorobacteno glucosil transferasa [EC: 2.-.-.-]	Q8KB11	(Maresca & Bryant, 2006)
Formación de carotenoides acíclicos	CrtC	Caroteno 1,2-hidratasa [EC:4.2.1.131]	P17058 *	(Hiseni, Arends, & Otten, 2011)
Formación de carotenoides acíclicos	CrtF	Dimetilesferoideno O-metiltransferasa [EC: 2.1.1.210]	P0CY89 *	(Scolnik, Walker, & Marrs, 1980)
Formación de carotenoides acíclicos	CrtD	1-hidroxycaroteno 3,4-desaturasa [EC: 1.3.99.27]	P17059 *	(Gerjets, Steiger, & Sandmann, 2009)
Formación de carotenoides acíclicos	CrtA	Esferoideno monooxigenasa [EC:1.14.15.9]	P17055 *	(Šlouf et al., 2012)

Síntesis de isoprenos

La síntesis de isoprenos (SI) se relaciona con la biosíntesis de carotenoides, debido a que durante la SI se producen los precursores necesarios para la formación de los carotenoides (**Figura 9**). Los precursores formados durante la SI son el isopentenil difosfato (IPP) y el dimetil difosfato (DMAPP) (Tsuchiya et al., 2005). El IPP y el DMAPP son moléculas de cinco carbonos que pueden ser formadas a través de las vías independientes del mevalonato (MVA) o de la deoxixilulosa-5-fosfato (DXP o MEP). La diferencia principal entre la vía del MVA y la vía del DXP, es que el inicio de la vía del MVA depende de la unión de dos moléculas de Acetil-CoA, mientras que en la vía DXP se requiere la unión de piruvato y D-gliceraldehído 3-fosfato (Lange, Rujan, Martin, & Croteau, 2000). En hongos, algas, plantas y animales, la formación de isoprenos depende de la vía MVA; en cambio, en bacterias, la formación de isoprenos ocurre generalmente por la vía DXP, excepto en la delta proteobacteria *Myxococcus fulvus* y en la Chloroflexi *Chloroflexus aurantiacus,* que dependen de la vía MVA (Lange et al., 2000; Rosa-Putra, Disch, Bravo, & Rohmer, 1998).

Una vez que se han formado el IPP y el DMAPP, se une una molécula de IPP con una molécula de DMAPP para formar geranil pirofosfato (GPP). Posteriormente, una molécula extra de IPP es adicionada al GPP para formar el compuesto farnesil difosfato (FPP). La enzima encargada de la formación del FPP es la geranilgeranil difosfato sintasa (CrtE) (Shivanand et al., 1992). A partir de la formación del FPP (C_{15}), pueden formarse apocarotenoides u homocarotenoides, dependiendo de la enzima que actúe en el substrato FPP. Si actúa la enzima diapofitoeno sintasa (CrtM), entonces se formarán apocarotenoides (< 40 átomos de carbono); por el contrario, si se presenta otra vez la enzima geranilgeranil difosfato sintasa (CrtE), se formarán homocarotenoides (≥ 40 átomos de carbono).

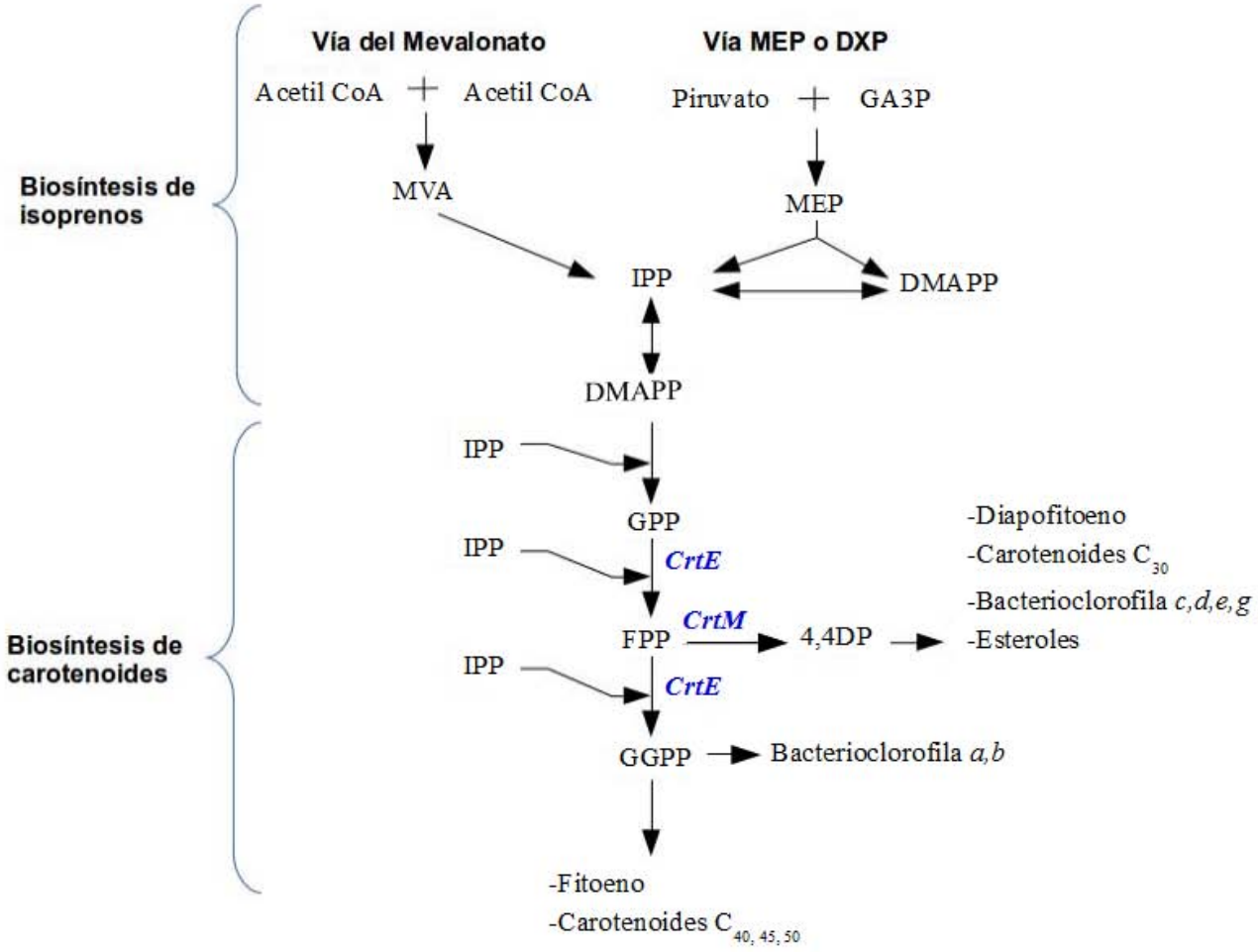


Figura 9. Biosíntesis de isoprenos y de carotenoides en bacterias. Abreviaciones: vía del mevalonato (MVA); 2C-metil-deritritol-4-fosfato (MEP); isopentenil difosfato (IPP); dimetilalil difosfato (DMAPP); geranil difosfato (GPP); farnesil difosfato (FPP); geranilgeranil difosfato (GGPP) (Imagen modificada de Armstrong, 1997).

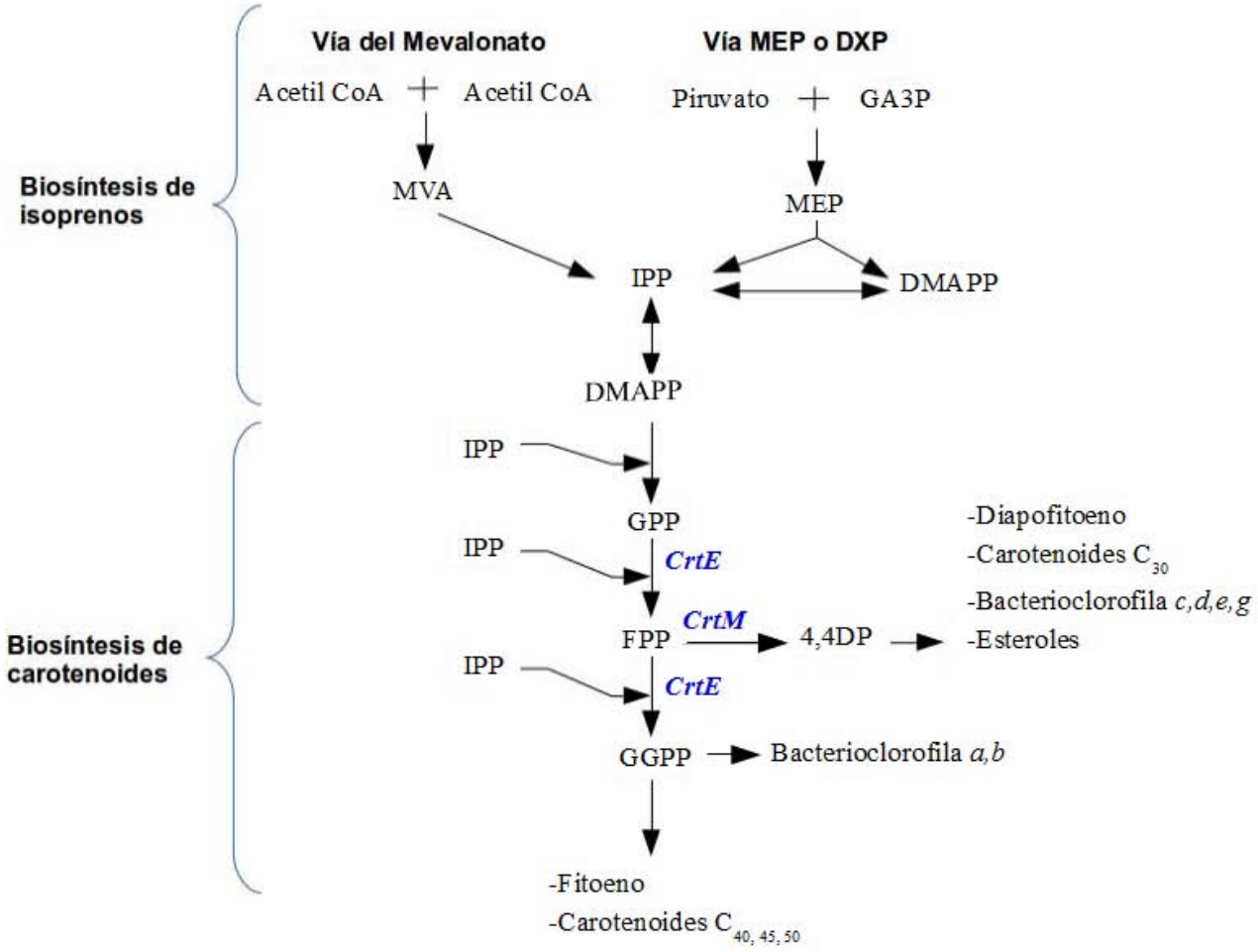
Síntesis de diapofitoeno

La síntesis de diapofitoeno comprende la formación de apocarotenoides a partir del FPP. En esta vía, la enzima diapofitoeno sintasa (CrtM), será la primera en reaccionar con el FPP al unir otra

molécula de FPP y formar prescualeno difosfato (C_{30}). Posteriormente, la diapofitoeno sintasa (CrtM) formará al 4'4-diapofitoeno que después será desaturado por la enzima 4,4'-diapofitoeno desaturasa (CrtN) para formar 4'4 neurosporeno (Furubayashi, Li, Katabami, Saito, & Umeno, 2014). Durante la vía de síntesis de diapofitoeno, se forma el apocarotenoide estafiloxantina en donde se requiere la participación sistemática de las enzimas diapolicopeno oxigenasa (CrtP2); 4'4-diaponeurosporenoato glicosil transferasa (CrtQ2); y glicosil-4,4'-diaporeurosporenoato aciltransferasa (CrtO2) (Pelz et al., 2005; T. D. Schneider & Stephens, 1990).

Síntesis de fitoeno

La formación de homocarotenoides comienza con la unión de una molécula de FPP con una de IPP para formar geranilgeranil difosfato (GGPP) (C_{20}), empleando a la enzima geranilgeranil difosfato sintasa (CrtE) para catalizar la reacción (Armstrong, 1997). Posterior a la formación de GGPP, se produce el precursor de todos los homocarotenoides, el prefitoeno difosfato (C_{40}) y posteriormente la formación de carotenoides continúa con la síntesis del prefitoeno. El prefitoeno difosfato es formado por la actividad de la enzima fitoeno sintasa, CrtB (AL2 en eucariontes), que también actúa en la reacción posterior para la formación de fitoeno (Hoshino et al., 1993). La proteína CrtB2 con actividad bifuncional de undecaprenil pirofosfato y fitoeno sintasa, también participa en la síntesis de prefitoeno a fitoeno, aunque adicionalmente cataliza la condensación de IPP (Heider et al., 2012).



Deshidrogenación del fitoeno

El fitoeno es deshidrogenado repetidas veces por diversas enzimas para formar licopeno. En cianobacterias, algas y plantas, la desaturación del fitoeno ocurre en cuatro desaturaciones que dependen de la actividad conjunta de dos enzimas, la 15-cis-fitoeno desaturasa CrtP (PDS en eucariontes) y la gamma-caroteno desaturasa, CrtQ (ZDS en eucariontes) (Liang et al., 2006). En cambio, en bacterias anoxigénicas fotosintéticas, bacterias no fotosintéticas y hongos, la desaturación

del fitoeno puede ocurrir en tres desaturaciones consecutivas para formar neurospoeno o en cuatro desaturaciones para formar licopeno (Armstrong, 1997). En algunos casos, como en *Gloebacter violaceus* PCC 7421, la enzima fitoeno desaturasa, tipo CrtI, puede realizar las cuatro desaturaciones sin ayuda de otra desaturasa (Liang et al., 2006; Takaichi, Maoka, & Masamoto, 2001). La deshidrogenación del fitoeno puede ocurrir mediante dos vías: la primera vía, comienza con la deshidrogenación del fitoeno a zeta-caroteno, por la actividad de CrtP y termina con la formación de licopeno al reaccionar la enzima CrtQ con el zeta caroteno. La segunda vía, consiste en la formación de licopeno a partir de la actividad de CrtP; zeta-caroteno isomerasa (Z-ISO); CrtQ y prolicopeno isomerasa (CrtH, CrtISO en eucariontes) (Phadwal, 2005). En la primer vía de deshidrogenación de fitoeno, existen enzimas que pueden actuar de forma alternativa para formar licopeno, como la fitoeno desaturasa CrtI, que actúa alternativamente en la formación de neurospoeno y licopeno (Maresca et al., 2007). La enzima zeta-caroteno desaturasa, CarA2, también puede actuar después de fitoeno y formar zeta-caroteno. De esta manera, la enzima all-trans-zeta-caroteno desaturasa, (CarC), puede llevar a cabo la formación de neurospoeno.

Ciclización de licopeno

La ciclización del licopeno ocurre en bacterias mediante ciclasas tipo epsilon o beta. Las ciclasas tipo epsilon añaden un anillo al substrato licopeno; en tanto que las ciclasas tipo beta añaden dos anillos, uno en cada extremo del licopeno (F. X. Cunningham et al., 1996). La epsilon-ciclasa LcyE, participa en las reacciones de formación de delta-caroteno y epsilon caroteno. A comparación de las enzimas tipo –epsilon ciclasas, las enzimas beta ciclasas son más diversas y actúan en reacciones relacionadas con la formación de gamma-caroteno, beta-caroteno, beta-zeacaroteno y 7,8-dihidro-beta-

caroteno. Las beta ciclasas conocidas son la CrtY; la ciclasa descrita en cianobacterias CrtL; y las enzimas licopeno ciclasas CruA y CruP, que sólo intervienen en la formación de gamma-caroteno y beta caroteno en bacterias fotosintéticas (Maresca et al., 2007). Otra beta-ciclasa descrita es la enzima bifuncional 15-cis-fitoeno sintasa/ licopeno beta ciclasa (AL2), la cual ha sido encontrada en hongos (Arrach et al., 2002). Las enzimas relacionadas con las reacciones posteriores a la ciclización de licopeno son beta-caroteno hidroxilasa (CrtR), que permite la formación de beta-criptoxantina; y la enzima beta-caroteno 3-hidroxilasa (CrtZ), que se relaciona con la formación de luteína, zeaxantina, beta-criptoxantina, adonixantina y foenicoxantina (Klassen, 2010). Para la formación de otros otros compuestos carotenoides como el isorenieraten, se requiere el sustrato beta caroteno y la actividad de isorenierateno sintasa (CrtU); para formar cantaxantina se necesita el beta caroteno cetolasa tipo CrtO y CrtW; y para la formación de zeaxantina diglucósido se requiere CrtX (Frigaard, Maresca, Yunker, Jones, & Bryant, 2004a; Misawa et al., 1990).

Formación de carotenoides acíclicos

Los carotenoides acíclicos se forman comúnmente durante la formación de carotenoides de 40 átomos de carbono cuando no actúan ciclasas. Las enzimas descritas que participan en la formación de carotenoides acíclicos son cuatro: la caroteno 1,2-hidratasa (CrtC); dimetilesferoideno O-metiltransferasa, (CrtF); 1-hidroxicaroteno 3,4-desaturasa, (CrtD); y esferoideno monooxigenasa (CrtA) (Badenhop, Steiger, Sandmann, & Sandmann, 2003; Umeno et al., 2005).

Antecedentes

Los carotenoides han sido ampliamente estudiados en plantas (F. X. Cunningham & Gantt, 1998; Hirschberg et al., 1997), algas (L. Barredo, 2012; Takaichi, 2011), hongos (Goodwin, 1980) y en algunos phyla de bacterias (Armstrong, 1997; Phadwal, 2005; Sieiro et al., 2003). En bacterias, los carotenoides han sido estudiados desde diversas perspectivas y áreas de trabajo como bioquímica (Britton, Armitt, Lau, Patel, & Shone, 1982) genética (Armstrong, 1997), biología molecular (Gómez-García & Ochoa-Alejo, 2013), evolución (Liang et al., 2006), ingeniería metabólica (Das et al., 2007), genómica (Da Silva-Mendes, Fontes-Soares, & Cardoso-Costa, 2015; Lohr, Im, & Grossman, 2005), entre otras áreas. En el área de genómica, se han realizado diversos estudios comparativos enfocados en la biosíntesis de carotenoides, siendo los más recientes los relacionados con la determinación de algunas de las proteínas de la BC en el phylum Cyanobacteria (Liang et al., 2006) y en otros phyla fotosintéticos (Klassen, 2010). Los carotenoides han sido descritos principalmente en cianobacterias (Liang et al., 2006; Tóth et al., 2015) y en otras bacterias fotosintéticas (D'Haene, Crouch, Jones, & Frese, 2014). Existe poca información sobre la presencia y función de los carotenoides en bacterias no fotosintéticas (Goodwin, 1972; Jeon, Kim, Jung, & Park, 2012; M M Mathews & Sistrom, 1959; Micheline M Mathews & Krinsky, 1965). Otros estudios de carotenoides en bacterias han sido enfocados principalmente en la caracterización de nuevos compuestos carotenoides (Hiseni et al., 2011; Maresca & Bryant, 2006; Steiger, Mazet, & Sandmann, 2003; Tao et al., 2005); en la descripción de su biosíntesis (Armstrong, 1997); en la relación que guardan con la fotosíntesis (Cogdell et al., 2000) y los roles que tienen los pigmentos carotenoides para la adecuación de las bacterias a sus respectivos ambientes (Cogdell et al., 2000; Dieser, Greenwood, & Foreman, 2010; Jáuregui-Carranco et al., 2011; Khaneja et al., 2010a; Paiva & Russell, 1999; Rottem, Gottfried, & Razin, 1968).

Objetivos

Objetivo general

Determinar los patrones de distribución de las proteínas de la biosíntesis de carotenoides en los principales phyla del dominio bacteria.

Objetivos particulares

1. Describir la abundancia de las secuencias homólogas de las proteínas de la BC en los phyla de bacterias analizados.

2. Evaluar la distribución de las secuencias homólogas de las proteínas de la BC en los phyla de bacterias estudiados.

Metodología

1. Selección de bacterias y obtención de su proteoma

Se seleccionaron 149 organismos de los phyla más estudiados del dominio Bacteria. Entre las bacterias elegidas están aquellas que utilizó Ciccarelli (*et al*., 2006), en la reconstrucción filogenética del árbol de la vida (ver **Anexo 1**). El criterio de selección utilizado por Ciccarelli (*et al*., 2006) en la reconstrucción filogenética del árbol de la vida permitió seleccionar bacterias con genoma totalmente secuenciado (hasta la fecha de la publicación del artículo), representantes de los phyla más estudiados del dominio Bacteria y que además compartieran 31 familias de genes ortólogos (Ciccarelli et al., 2006).

Se seleccionaron los proteomas de la base de datos del NCBI *Entrez Genome*, en el servidor FTP (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria). Se descargaron únicamente proteomas completos, dando prioridad a los proteomas usados por Ciccarelli (*et al*., 2006) y en caso de encontrar más de un proyecto de secuenciación, se descargó el proteoma de referencia del servidor de NCBI (http://www.ncbi.nlm.nih.gov/refseq/). El proteoma de las bacterias fue descargado (Marzo 2015) en archivos de texto con extensión .faa. En los casos en donde se encontraron más de un archivo .faa (*ie*. secuencias de proteínas de cromosoma y plásmidos), se descargaron ambos archivos y se concatenaron en un solo archivo de texto. El proteoma fue descargado utilizando el comando "**wget**". El paquete wget permite descargar archivos usando protocolos de internet como FTP. Un ejemplo del uso del comando wget se muestra a continuación:

```
$ wget ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/Acidobacterium_capsulatum.faa/*.faa

> Acidobacterium_capsulatum.faa
```

Para concatenar los archivos de texto .faa en un solo archivo, se uso el comando "**cat**". Este comando permite unir varios archivos de texto en un solo archivo de salida. Un ejemplo del uso del comando cat, se muestra a continuación:

```
$ cat NC_004320.faa NC_004319.faa > Corynebacterium_efficiens.faa. conct
```

en donde:

NC_004320.faa y NC_004319.faa: corresponden a los archivos de texto con extensión .faa que fueron encontrados para la misma bacteria.

2. Selección de las secuencias de proteínas de la biosíntesis de carotenoides en bacterias

Para obtener un compendio de las secuencias de proteínas que participan directamente en las reacciones de la biosíntesis de carotenoides (BC) en bacterias, se llevó a cabo una revisión bibliográfica en el servidor PubMed (http://www.ncbi.nlm.nih.gov/pubmed). Paralelamente, se buscaron secuencias de aminoácidos de la BC en bases de datos de rutas metabólicas como KEGG (http://www.kegg.jp/), MetaCyc (http://metacyc.org/) y ProCarDB (http://bioinfo.imtech.res.in/servers/procardb). Se enlistaron inicialmente 37 proteínas que participan en la BC de plantas, algas y bacterias, de las cuales sólo se conservaron 25 proteínas que participan directamente en la BC de bacterias.

El criterio de selección de las secuencias se basó en las pautas propuestas por Klassen (2010). Dichas pautas son las siguientes: proteínas que participen en la BC y que su actividad enzimática haya sido descrita por estudios bioquímicos *in vitro*; proteínas de bacterias en las que la expresión recombinante de sus genes codificantes (en un hospedero no carotenogénico) haya resultado en una reacción anabólica con función en la BC; proteínas de bacterias en las que la mutación de algunos de sus genes codificantes resulte en la pérdida de función en la BC. Adicionalmente, se consideraron proteínas de la BC que estuvieran anotadas en la base de datos UniProtKB/SwissProt *(ie.* CrtB2, proteína bifuncional isoprenil transferasa / fitoeno sintasa [EC:2.5.1.32] / [EC= 2.5.1.-]) y en UniProtKB/TrEMBL *(ie.* Crup, licopeno ciclasa [EC: 5.5.1.19]; CrtR,Beta-caroteno hidroxilasa [EC:1.14.13.-]; CrtU, Isorenierateno sintasa [EC:1.-.-.-]; CruC, Clorobacteno glucosil transferasa [EC: 2.-.-.-]) (Apweiler et al., 2004). La relación de las 25 proteínas de la BC usadas, se muestra en la **Tabla 1**.

2.1 Representación gráfica de la biosíntesis de carotenoides con las proteínas seleccionadas

Las 25 proteínas seleccionadas previamente, fueron agrupadas en las reacciones principales de la biosíntesis de carotenoides: biosíntesis del isopreno, síntesis de diapofitoeno, síntesis de fitoeno, deshidrogenación del fitoeno, ciclización del licopeno, reacciones posteriores a la ciclización de licopeno y formación de carotenoides acíclicos (Klassen, 2010; Phadwal, 2005). En la **Tabla 1** se muestra la organización de las proteínas en las diferentes reacciones. Con la finalidad de representar gráficamente la participación de las 25 proteínas en la BC, se generó un diagrama en el programa PathVisio, versión 3.1.3. (van Iersel et al., 2008). En el programa PathVisio, se trazaron las principales reacciones de la BC y se ubicó cada una de las 25 proteínas seleccionadas en sus reacciones correspondientes.

3. Entrenamiento de perfiles de Modelos Ocultos de Márkov y búsqueda de homólogos

3.1 Obtención de secuencias semilla para el entrenamiento de perfiles HMM

Las secuencias de referencia de las 25 proteínas de la BC (**Tabla 1**), fueron obtenidas de la base de datos de UniProtKB/SwissProt. Posterior a la identificación de las secuencias de referencia, se buscaron sus homólogos (parálogos y ortólogos) en la base de datos de UniProtKB y se delimitó la búsqueda para hallar únicamente proteínas presentes en bacterias. Esta búsqueda de homólogos se realizó con el programa BLASTP (e-value $1x10^{-6}$, tamaño de palabra 10), versión 2.2.29+, en el

servidor de UniProtKB (http://www.uniprot.org/blast/) (Shiryev, Papadopoulos, Schäffer, & Agarwala, 2007). Para cada una de las búsquedas, se eligieron 250 resultados que tuvieran al menos 30% de identidad y que representaran a más de un phylum. Las secuencias homólogas fueron descargadas en archivos individuales, tipo FASTA.

3.2 Alineamiento múltiple de secuencias

Las secuencias homólogas de cada una de las proteínas de la BC, fueron alineadas con el programa de alineamiento múltiple de secuencias MAFFT, versión 7 (Katoh & Standley, 2013). El programa MAFFT y la curación manual de los alineamientos múltiples de secuencias (AMS) fueron ejecutados en la plataforma del programa Jalview, (versión 2.8.1) (Martin, Procter, Waterhouse, Shehata, & Barton, 2013). Jalview es un programa que permite visualizar, analizar y editar alineamientos múltiples de secuencias (AMS). La curación manual de los AMS consistió en eliminar las filas de secuencias que no presentaran los motivos conservados por la familia de proteínas. Los AMS curados fueron descargados en archivos de texto individuales y se les asignó la terminación ".aln".

3.3 Entrenamiento de los Modelos Ocultos de Márkov

Se entrenaron 25 perfiles HMM a partir de los AMS curados previamente (ver Metodología, Alineamiento múltiple de secuencias). Los perfiles HMM fueron entrenados en el programa HMMER, versión 3.1b1 (Finn et al., 2011; Katoh & Standley, 2013). El comando usado para construir perfiles HMM fue "**hmmbuild**", el cual permite generar perfiles HMM a partir de AMS. Los perfiles HMM fueron nombrados con el nombre de la proteína a la que representaron, junto con la terminación "hmm". El comando hmmbuild fue usado de la siguiente manera:

```
$ hmmbuild CrtP2.hmm CrtP2.aln
```

en donde:

CrtP2.hmm: es el nombre del archivo que contiene el perfil HMM

CrtP.aln: es el archivo que contiene al AMS curado

3.4 Evaluación de perfiles HMM

Primera evaluación: Capacidad de los perfiles HMM para identificar proteínas homólogas en proteomas control

Antes de realizar las búsquedas de los perfiles HMM en los proteomas de bacterias, se realizaron diversas pruebas. La primer prueba consistió en evaluar la capacidad de los perfiles para encontrar proteínas homólogas en algunos proteomas control (20 proteomas en donde anteriormente se ha reportado la presencia de 25 proteínas seleccionadas). La lista de los 20 proteomas analizados y las proteínas que deberían encontrarse en cada proteoma, se muestran en el **Anexo 2**. La búsqueda de perfiles fue realizada con ayuda del programa HMMER. El comando utilizado para buscar secuencias homólogas en los proteomas fue "**hmmsearch**". Este comando fue usado de la siguiente manera:

```
$ hmmsearch --tblout CrtBSynechocystis.out CrtB.hmm /Directorio/bacterias/Synechocystis
```

en donde:

hmmsearch: es la instrucción para buscar un perfil HMM en una base de datos

--tblout: es una opción del comando hmmsearch que permite obtener resultados en formato tabular

CrtBSynechocystis.out: define el nombre del archivo de texto con los principales resultados de la búsqueda en el proteoma de *Synechocystis*

CrtB.hmm: es el perfil HMM que será buscado, en este caso el de la proteína CrtE

/Directorio/bacterias/Synechocystis: es la ruta en la que se encuentra el proteoma de prueba que será analizado

A continuación se muestra un fragmento de un archivo de salida obtenido por el comando hmmsearch y en amarillo se resaltan las proteínas homólogas que fueron encontradas en una de las búsquedas. En la primer proteína hallada phytoene synthase, el valor de 2.5e-81 corresponde al *E-value* y 270.5 al *bit score.*

```
#                                                                    --- full sequence ---- ---
best 1 domain ----

# target name          accession  query name           accession    E-value  score  bias    description
of target


gi|383322181|ref|YP_005383034.1| -          crtB                 -            2.5e-81  270.5   0.0
phytoene synthase [Synechocystis sp. PCC 6803 substr. GT-I]

gi|383323764|ref|YP_005384618.1| -          crtB                 -            3.7e-18   63.3   0.1
hypothetical protein SYNGTI_2856 [Synechocystis sp. PCC 6803 substr. GT-I]

#

...
```

Segunda evaluación: Capacidad de los perfiles HMM para discriminar entre proteínas homólogas

La segunda prueba consistió en analizar la capacidad de los perfiles para discriminar entre proteínas anotadas como homólogas y aquéllas que no lo eran. En este caso, se generó una base de datos que contuviera 250 secuencias homólogas de cada perfil, más otras 6 000 secuencias no homólogas. En total, la base de datos contenía 6 250 secuencias de proteínas. La base de datos fue llamada megafasta.txt. Con ayuda del programa HMMER, se buscaron las 250 secuencias homólogas de cada perfil, en la base de datos megafasta. El comando usado fue el siguiente:

```
$ hmmsearch crtA.hmm megafasta.txt > crtA.com
```

en este caso crtA.hmm es el perfil buscado en la base de datos megafasta.txt. El archivo generado con las proteínas homólogas halladas es crtA.com.

Cuando se observó que algunos perfiles encontraban falsos positivos, se analizó la probabilidad de que los perfiles pudieran tener secuencias no homólogas. Los falsos positivos fueron considerados como aquellas proteínas que tuvieran fragmentos de secuencia similares a los que buscaba el perfil HMM, pero que no conservaran los motivos que caracterizaban a la familia de proteínas en cuestión. Para evitar la selección de falsos positivos, se definió para cada perfil un rango de selección con la finalidad de discriminar entre secuencias homólogas y no homólogas, a partir de su *bit score.* El *bit score* fue elegido debido a que su cálculo no depende del tamaño de la base de datos, y por lo tanto, permite comparar la similitud entre secuencias obtenidas de diferentes proteomas (Eddy, 2010). El

rango de selección fue definido con el *bit score* máximo y mínimo posible, que pudieran tener las proteínas homólogas conocidas para dicha familia de proteínas. El *bit score* máximo y mínimo se obtuvo al comparar el perfil HMM de cada proteína contra cada una de sus 250 secuencias con las que fueron calibrados. Para analizar detenidamente los datos que definieron al rango de selección, se decidió graficar la distribución de los *bit score*, en un diagrama de caja, el cual fue generado en el programa R (versión 3.0.2) (R Core Team, 2013a). Los diagramas de caja se muestran en el **Anexo 5**. El rango de selección, la cantidad de secuencias con las que fue calibrado cada perfil y la longitud de cada perfil, se muestra en el **Anexo 2**.

3.5 Representación gráfica de los perfiles HMM

Para representar gráficamente los perfiles HMM, se generaron HMM logos con el programa *Skylign* (http://skylign.org/). Los HMM logos fueron descargados como archivos de imagen y almacenados para análisis posteriores. Los HMM logos se muestran en el **Anexo 6**.

3.6 Búsqueda de perfiles HMM en proteomas de bacterias

Los 25 perfiles HMM generados previamente, fueron usados para buscar secuencias homólogas en los 147 proteomas seleccionados. Se usó el programa HMMER, con su comando de búsqueda "**hmmsearch**". El uso del comando hmmsearch ya ha sido descrito anteriormente (ver Evaluación de perfiles HMM). En este caso, la base de datos en donde se realizaron las búsquedas con los perfiles HMM, fueron los 147 proteomas. Un ejemplo del comando usado para buscar el perfil CrtB en todos los proteomas, se muestra a continuación.

```
$ hmmsearch --tblout CrtB(nombre_del_proteoma).out CrtB.hmm

/Directorio/bacterias/Proteomas
```

En este caso, la opción --tblout. permite obtener archivos de texto, en formato tabular, con los mejores aciertos para el perfil CrtB.hmm. El proteoma en el que se buscó CrtB.hmm, se localiza en el directorio mencionado (Directorio/bacterias/Proteomas). En la carpeta /Proteomas, se encuentran los 147 archivos de texto que contienen los proteomas de las bacterias analizadas. Para cada proteoma, se obtuvieron 25 archivos de texto que contenían los mejores aciertos para cada perfil. A continuación, se muestra un archivo de texto generado durante la búsqueda del perfil CrtM, en el proteoma de *Staphylococcus aureus*, subespecie *aureus* Mu50:

```
.                                        E-value   bit score     bias

gi|15925552|ref|NP_373086.1| -   crtM     8.5e-207     682.4     20.6 squalene desaturase [Staphylococcus
aureus subsp. aureus Mu50]
```

4. Procesamiento de datos

A partir de los 3,675 archivos de texto generados, se tomaron en consideración sólo a las proteínas que se encontraban dentro del rango de selección de los perfiles, para lo cual se empleó el lenguaje de programación para procesamiento de textos "**awk**". Para entender mejor el funcionamiento de awk, se muestra a continuación un archivo de salida antes del procesamiento:

```
 #                                                                          --- full sequence ---- ---
best 1 domain ---- --- domain number estimation ----
Bit score
```

```
# target name                 accession  query name           accession    E-value  score  bias   E-
value  score  bias   exp reg clu  ov env dom rep inc description of target

Proteína aceptada

#        ------------------- ---------- -------------------- ---------- --------- ------ -----
--------- ------ -----   --- --- --- --- --- --- --- --- ---------------------gi|47529696|ref|
YP_021045.1| -          crtE                 -           8.8e-110  364.8   1.4  9.7e-110  364.6   1.4
1.0   1   0   0   1   1   1   1 geranyltranstransferase [Bacillus anthracis str. 'Ames Ancestor']

gi|47526809|ref|YP_018158.1| -          crtE                 -           2.6e-46  156.4   0.2   3.7e-
46  155.9   0.2   1.3   1   0   0   1   1   1   1 heptaprenyl diphosphate synthase component II
[Bacillus anthracis str. 'Ames Ancestor']Proteína no aceptada

## Program:         hmmsearch

# Version:         3.1b1 (May 2013)

# Pipeline mode:   SEARCH

# Query file:      crtE.aln.hmm

# Target file:
/home/pako/crtlitbase/blakeggmsa/mafftWS2/hmmbuild/hmmlogo/hmmsearchout/Bacillus_anthracis__Ames_Ancest
or__uid58083/Bacillus_anthracis__Ames_Ancestor__uid58083.faa.conct

# Option settings: hmmsearch --tblout crtEBacillus_anthracis__Ames_Ancestor__uid58083.out
crtE.aln.hmm
/home/pako/crtlitbase/blakeggmsa/mafftWS2/hmmbuild/hmmlogo/hmmsearchout/Bacillus_anthracis__Ames_Ancest
or__uid58083/Bacillus_anthracis__Ames_Ancestor__uid58083.faa.conct

# Current dir:     /home/pako/db/alnm/hmm

# Date:            Tue Sep  2 12:56:57 2014
```

El comando awk fue usado para identificar la columna de *bit score* y en dicha columna, seleccionar sólo las filas que estuvieran dentro del rango de aceptación de los perfiles HMM. En las filas seleccionadas, se le pidió a awk imprimir los renglones que contuvieran información relevante sobre el *bit score*, el nombre de la proteína y el organismo en el que se encontró la proteína. La siguiente línea muestra su funcionamiento:

```
awk '{if($6 >= 285) print $6, $1, $19, $20, $21, $22, $23, $24, $25, $26, $27, $28}'
/home/genom/Bacillus_anthracis/crtE.out | grep -e '^[0-9][0-9][0-9].' > crtE.fin
```

en donde:

awk: es una herramienta de UNIX que permite buscar patrones y si los encuentra, aplicar acciones como imprimir los resultados en la terminal o en un archivo de texto.

La sección '{if($6 >= 285) es una función condicional, que de ser verdadera, se completará con la acción print. El condicional if le dice al programa awk, que si encuentra un valor en la columna seis ($6) (perteneciente al bit score), mayor o igual a 285 (>= 285), entonces imprima las columnas 6, 1 y de la 19 a la 28 (print $6, $1, $19, $20, $21, $22, $23, $24, $25, $26, $27, $28}'). Las columnas 1, 19-28, contienen información de las proteínas halladas dentro del rango de selección. La instrucción de búsqueda está enfocada en este caso en el archivo crtE.out (/home/genom/Baillus_anthracis/crtE.out). Una vez que se obtuvieron los segmentos de texto definidos, se redirigió la salida a grep. El comando grep permitió seleccionar sólo las filas que comenzaran con tres números seguidos y un punto decimal e imprimirlos en un archivo de texto. En este ejemplo el archivo de texto es llamado crtE.fin. Para cada proteoma, se generaron archivos de texto ".fin" correspondientes a las proteínas homólogas halladas por los perfiles HMM. Un archivo de texto ".fin" producido después del procesamiento con awk se muestra a continuación :

```
364.8 gi|47529696|ref|YP_021045.1| geranyltranstransferase [Bacillus anthracis str. 'Ames Ancestor']
```

4.1 Obtención de matriz de datos con los homólogos de las proteínas de la biosíntesis de carotenoides

Para analizar la abundancia y la distribución de las proteínas homólogas encontradas con los

perfiles HMM, se generó una matriz de datos. En las columnas se ordenaron las proteínas analizadas y en las filas se colocaron los nombres de los proteomas organizados por phylum. En las celdas intermedias se enlistaron las proteínas halladas para cada proteoma. Al observar que existían proteomas con más de una copia para una misma proteína (probables proteínas parálogas), se cuantificó su abundancia. Los valores de abundancia posibles para un solo proteoma fueron definidos entre cero y cuatro proteínas homólogas (parálogas). La matriz con la abundancia de las proteínas en cada uno de los proteomas se muestra en el **Anexo 3**. La matriz generada fue usada posteriormente para los análisis de abundancia y distribución filogenética de proteínas de la BC.

5. Obtención de la abundancia de homólogos de las proteínas de la biosíntesis de carotenoides

La abundancia de las secuencias homólogas de la BC fue analizada en diferentes niveles: como abundancia total, a nivel de phylum y a nivel de proteoma.

5.1. Abundancia total

La abundancia total se obtuvo al contar todos los homólogos encontrados (**Anexo 3**). Adicionalmente, se reportó la cantidad de homólogos encontrados por cada una de las proteínas de la BC analizadas. Dicha abundancia fue graficada en un histograma, con ayuda del programa R (R Core Team, 2013b). A partir de ésta abundancia se obtuvo el porcentaje de cada proteína, tomando como el 100% el total de proteínas encontradas. El porcentaje fue representado en gráfico de pastel.

Posteriormente, se representó el porcentaje de cada proteína de la BC, en una imagen de la misma ruta (generada con el programa *PathVisio*). Esto, con la finalidad de comparar la importancia de cada proteína a partir de la proporción de su abundancia de homólogos (van Iersel et al., 2008).

5.2. Abundancia a nivel de phylum

En la abundancia a nivel de phylum se obtuvieron tres datos. El primer dato hace referencia a la abundancia de las proteínas por cada phylum (ver **Tabla 2**). El segundo dato deriva de la abundancia de cada phylum, normalizada entre el número de organismos analizados en el respectivo phylum (ver **Tabla 2**). El tercer dato, muestra la proporción en la que fueron encontradas las proteínas de la BC en cada phylum. Es decir, en el caso del phylum Cyanobacteria, se analizó la abundancia relativa de la proteína CrtE respecto al total de proteínas encontradas en cianobacterias.

5.3. Abundancia a nivel de proteoma

La abundancia a nivel de proteoma fue descrita con dos valores: el primer valor correspondió a la abundancia total de proteínas encontradas en el proteoma y el segundo valor correspondió a la abundancia relativa de las proteínas de la BC respecto al total de proteínas en el proteoma. Para calcular el total de proteínas predichas en cada proteoma, se realizó un conteo con el comando **"grep -c"** en cada uno de los archivos que contenían a los proteomas.

```
$ grep -c ">" /Directorio/proteoma
```

La lista con la cantidad de proteínas predichas en los proteomas, se muestra en el **Anexo 4**. La

abundancia relativa de las proteínas de la BC, fue calculada al dividir la abundancia de las proteínas de la BC, entre el total de proteínas cuantificadas en el proteoma (ver **Anexo 4**). La abundancia relativa fue graficada en un *heatmap* con el programa R con los valores normalizados por filas (R Core Team, 2013b). Las instrucciones usadas en R, para graficar el *heatmap* fueron las siguientes:

```
>gen<-read.csv("frec_en_genome.csv")
>row.names(gen)<-gen$Species_iTOL
>gen<-gen[,2:34]
>gen_matrix<-data.matrix(gen)
>mipaleta <- brewer.pal(9,"blue")
>pdf("heat_frec_genom.pdf", pointsize=10)
>gen_heatmap <- heatmap(gen_matrix, Rowv=NA, Colv=NA, cexRow=0.5, revC='true', ColSideColors, RowSideColors, scale="Row", trace,
margins=c(5,30))
>dev.off()
```

La primera fila crea la variable gen a partir del archivo de texto "frec_en_genomes.csv". La segunda instrucción row.names, permite asignar nombres a las filas, con los nombres descritos en la columna Species_iTOL (gen$Species_iTOL).

gen<-gen[,2:34]:señala que la variable "gen", será tratada como una matriz desde la columna 2 hasta la 34.

gen_matrix<-data.matrix(gen): permite convertir los valores seleccionados a una matriz denominada gen_matrix

mipaleta <- brewer.pal(9,"blue"): define los colores que serán usados en el heatmap dentro de la variable "mipaleta"

pdf("heat_frec_genom.pdf", pointsize=10):crea un pdf llamado “heat_frec_genom.pdf” con tamaño de 10 de puntero.

gen_heatmap <- heatmap(gen_matrix, Rowv=NA, Colv=NA, cexRow=0.5, revC='true', ColSideColors, RowSideColors, scale="none", trace, margins=c(5,30)): crea el heatmap y define las variables para que sea graficado.

dev.off(): fija el final del entorno para la creación del pdf.

6. Determinación de la distribución filogenética de los homólogos de las proteínas de la biosíntesis de carotenoides

Para analizar los patrones de distribución de las proteínas de la BC, se graficó un *heatmap* asociado a una filogenia. El *heatmap* fue construido a partir de la matriz de datos generada previamente (**Anexo 4**), en el programa R (R Core Team, 2013b). El *heatmap* de abundancias fue generado de forma equivalente a como se ha generado el *heatmap* de abundancias relativas (ver Abundancia a nivel proteoma). En este caso, se usó la base de matriz de datos del **Anexo 4** con la abundancia total de las proteínas de la BC. A continuación se muestran los comandos usados en R:

```
>gena<-read.csv("anexo4.csv")
>row.names(gen)<-gen$Species_iTOL
>gen<-gen[,2:34]
>gen_matrix<-data.matrix(gen)
>mipaleta <- brewer.pal(9,"blue")
>pdf("figura8.pdf", pointsize=10)
>gen_heatmap <- heatmap(gen_matrix, Rowv=NA, Colv=NA, cexRow=0.5, revC='true', ColSideColors,
```

```
RowSideColors, scale="None", trace, margins=c(5,30))
>dev.off()
```

La filogenia de Ciccarelli (**Figura 6**) fue acoplada al *heatmap* generado por R, haciendo coincidir el orden de la tabla con el de la filogenia. La edición de la figura se llevó a cabo con el programa de edición de imágenes InkScape (Owens, 2015). La filogenia propuesta por Ciccarelli (*et al*., 2006) fue modificada y obtenida del servidor iTOL (http://itol.embl.de/itol.cgi), en la sección de "*TREE OF LIFE*". La filogenia asociada a los *heatmap* de abundancia relativa a nivel proteoma y abundancia total, es la misma.

Resultados y discusión

Abundancia y distribución de las proteínas homologas

En total, se lograron identificar 253 secuencias homólogas en las bacterias analizadas (**Figura 10**). En este sentido, la abundancia total no puedo ser comparada con estudios previos debido a que existen diferencias entre las proteínas (Klassen, 2010), los phyla y la cantidad de proteomas analizados (Liang et al., 2006). En el **Anexo 3** se muestran la matriz de datos con las proteínas encontradas.

Las 253 proteínas homólogas se encontraron en 135 proteomas de bacterias. Lo anterior equivale a que el 93 % (135/149) de las bacterias analizadas presentaron al menos una proteína de la BC. En términos biológicos, esto significa que las proteínas de la BC han sido conservadas de forma heterogénea en el dominio Bacteria (L. Barredo, 2012), lo cual impide que los patrones de distribución de carotenoides puedan ser usados para clasificación taxonómica en bacterias (Takaichi, 2009).

Síntesis de isoprenos

Geranilgeranil difostato sintasa (CrtE) [EC=2.5.1.29]: Se observó que la abundancia de homólogos fue diferente dependiendo de la proteína de la BC analizada. En este sentido, la proteína geranilgeranil difostato sintasa (CrtE) fue la más abundante con más de 90 secuencias homólogas (**Figura 10),** lo que equivalió a que el 37% de los homólogos encontrados correspondieron a dicha proteína (**Figura 11**). Cabe recordar que la proteína CrtE participa en la formación de la estructura química primaria de los carotenoides, al unir la síntesis de isoprenoides y la formación del GGPP (Shivanand et al., 1992). La abundancia de la proteína CrtE confirma la hipótesis propuesta por Armstron (*et al.*, 1990), en la cual menciona que las enzimas que participan en las reacciones iniciales de rutas biosintéticas se encuentran

más conservadas a lo largo de la divergencia evolutiva de las bacterias (**Figura 12**). Estos mismos resultados han sido confirmados en estudios previos, como el de Liang (*et al*., 2006), en el cual menciona que la proteína CrtE fue una de las más conservadas de la BC en el phylum Cyanobacteria. Una de las explicaciones que ha sugerido Umeno (2005) sobre la conservación de la proteína CrtE es que las reacciones conservadas *upstream* o iniciales facilitan la síntesis del esqueleto principal de los carotenoides; lo que le da soporte a las reacciones siguientes que incrementan la diversidad y la especificidad de compuestos con determinadas especies.

En la **Figura 15**, se observó que algunos proteomas tenían ausente la proteína de CrtE y contaban con otras proteínas de la BC (ie. *Chlorobium tepidum*, *Treponema dentícola*, *Mycobacterium* sp., *Corynebacteriumsp* sp., *Streptomyces coelicolor*, *Termotoga maritima* y *Deinococcus radiodurans*). Se consideró que la ausencia de la proteína CrtE pudo deberse a un error metodológico relacionado con los rangos de selección; sin embargo, se observó que la proteína CrtE jamás fue identificada. En sentido, Armstrong (1997) sugiere la hipótesis sobre la presencia de más de una GGPP sintasa en bacterias carotenogénicas para la acumulación alternativa de pigmentos, lo cual podría indicar que en este estudio podría haberse observado la presencia de una sola GGPP sintasa. Por lo anterior, se sugiere que para futuros estudios se contemple la presencia de más de una GGPP sintasa y evaluar su distribución en bacterias.

Contrario a lo anterior, se observó que algunas bacterias eran capaces de tener mas de una copia del gen que codifica a la proteína CrtE (ie. *Photorhabdus luminescens*, *Vibrio cholerae*, *Yersinia pestis Medievalis*). En el caso de *P. luminescens*, se sabe que es una enterobacteria bioluminescente, patógena de insectos y que aunque se ha predicho anteriormente la presencia algunos genes que codifican a proteínas de la BC, aún se desconoce su ruta metabólica (Duchaud et al., 2003; Gaudriault et al., 2006). Sobre *Yersinia pestis Medievalis* y *V. cholerae*, no se cuenta con información sobre biosíntesis de

carotenoides. Por lo anterior, se podría considerar que el perfil de la proteína CrtE podría haber reconocido homólogos muy remótos o falsos positivos.

La proteína CrtE se encontró distribuida en los phyla Firmicutes, Fibrobacteres, Planctomycetes, Spirochaetes, Actinobacteria, Fusobacteria, Cyanobacteria, Acidobacteria y en las clases Deltaproteobacteria, Alphaproteobacteria, Betaproteobacteria y Gammaproteobacteria (**Figura 16**). Al comparar la distribución de la proteína CrtE con reportes de estudios previos, se confirmó la presencia de la proteína CrtE en los phyla Cyanobacteria y Proteobacteria (Liang et al., 2006; Phadwal, 2005). A partir de la distribución y abundancia de la proteína CrtE, se dedujo que la proteína CrtE es una de las más conservadas a lo largo de los distintos phyla evaluados.

Dentro de las proteínas encontradas, se observó que existían proteínas que se encontraban fuera del rango de selección (285 a 391 bit score) del perfil HMM (ej. componente II de heptaprenil difosfato sintasa con bit score de 156.4. Lo anterior sugiere que se pudieron omitir proteínas pertenecientes a una familia de proteínas por la implementación de un rango de aceptación.

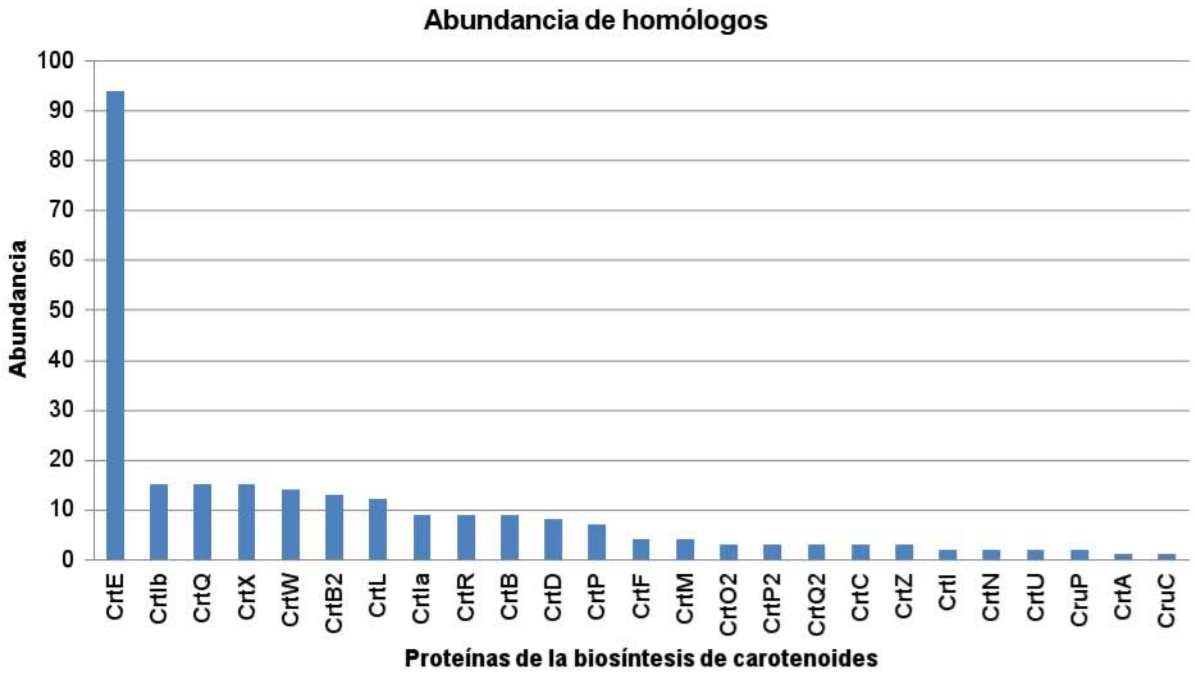


Figura 10. Abundancia de las proteínas de la biosíntesis de carotenoides en las bacterias analizadas. Las proteínas se encuentran organizadas de mayor a menor abundancia. La proteína CrtE (geranilgeranil difosfato sintasa), responsable de las primeras reacciones de la biosíntesis, fue la más abundante.

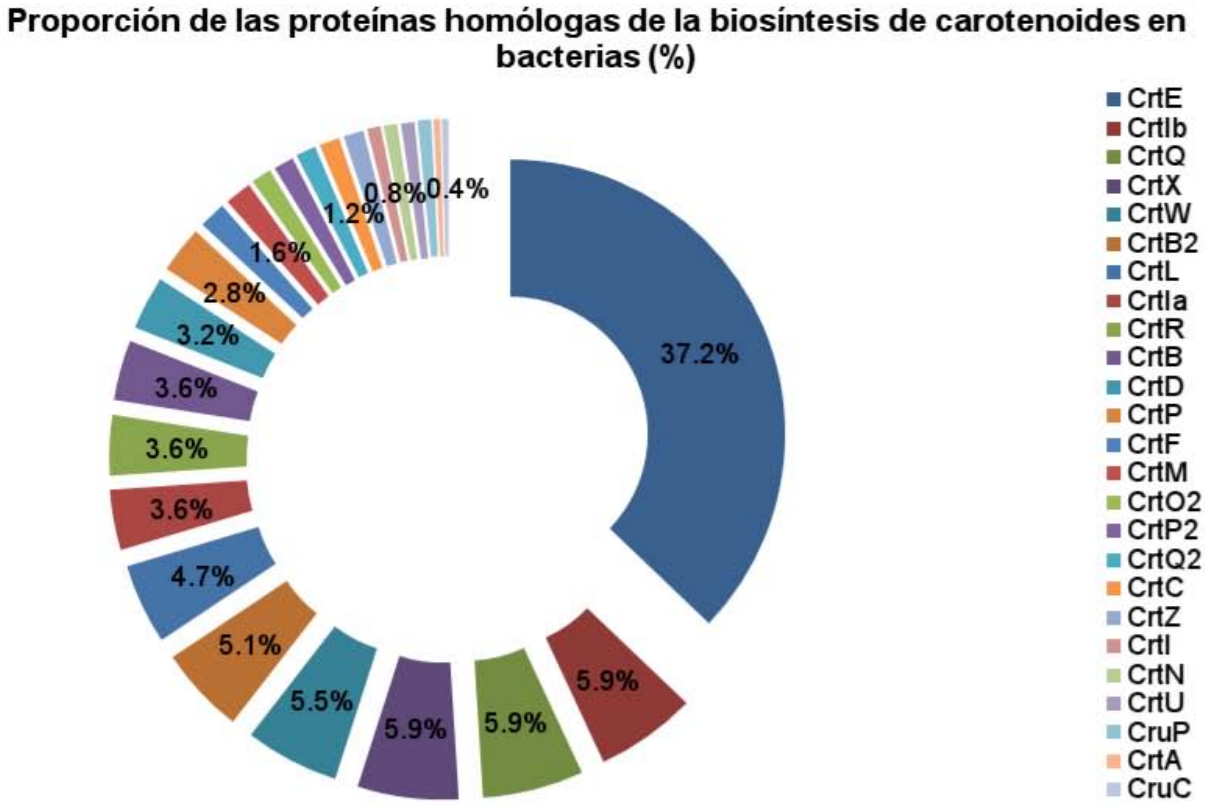


Figura 11. Porcentaje de la abundancia de las proteínas de la biosíntesis de carotenoides, considerando un total de 253 proteínas. Se observa que la proporción de CrtE es la mayor y el resto de las proteínas ocupa una proporción menor al 6%.

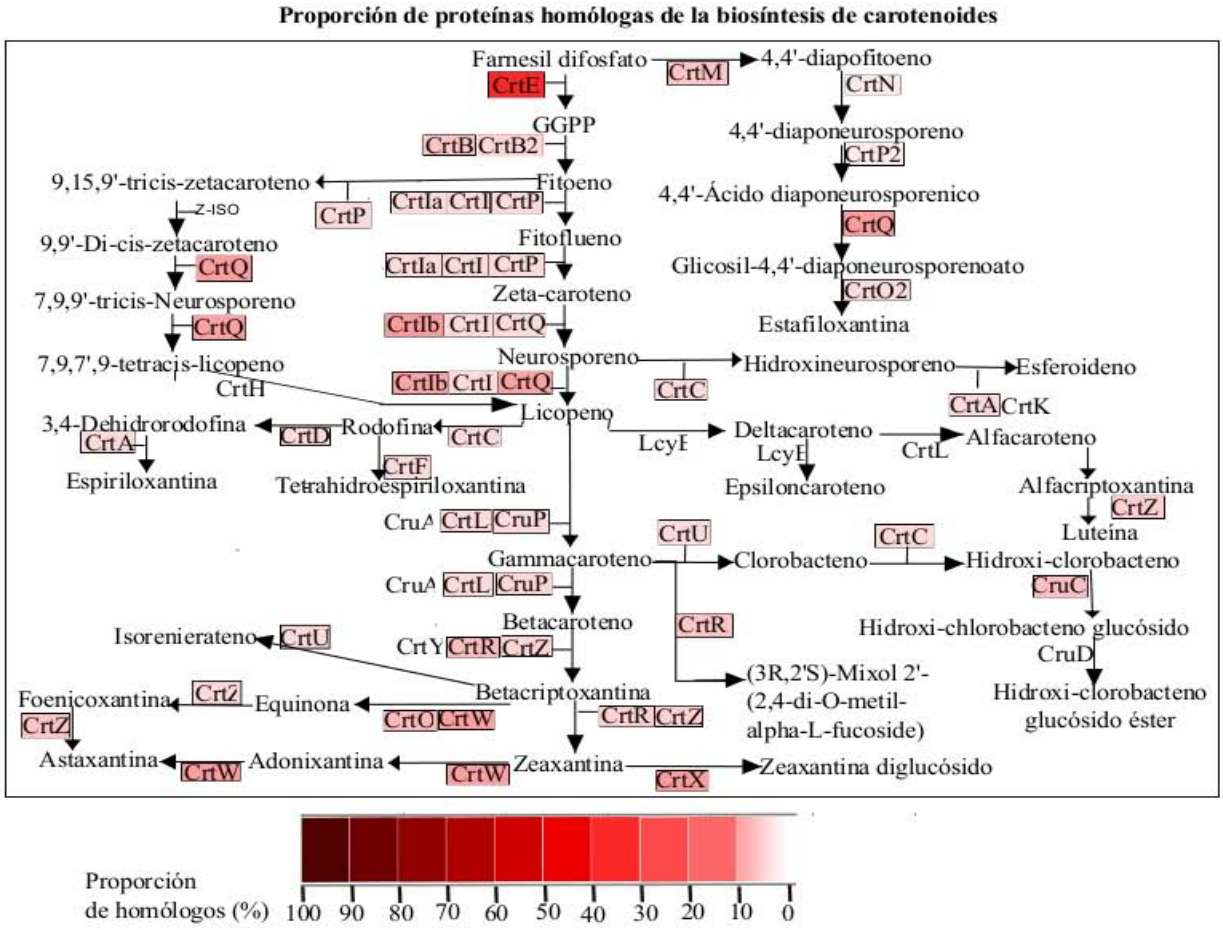


Figura 12. Representación de la abundancia de las proteínas en la ruta de BC. Las proteínas analizadas se muestran en rectángulos rojos. A mayor abundancia, el color rojo es más fuerte. Los metabolitos se muestran relacionados con flechas. La proteína CrtE, relacionada con la síntesis de precursores de la BC, se muestra en mayor contraste con 37%.

Síntesis de diapofitoeno

Dehidroscualeno sintasa (CrtM) [EC=2.5.1.96]; dehidroescualeno desaturasa (CrtN) [EC=1.3.8.2]; diapolicopeno oxigenasa (CrtP2) [EC=1.14.99.44]; 4'4-diaponeurosporenoato

glicosiltransferasa (CrtQ2) [EC=2.4.1.-]; Glicosil-4,4'-diaponeurosporeonato aciltransferasa (CrtO2) [EC=2.3.1.-]:: La abundancia de proteínas homólogas de CrtM, CrtN, CrtP2, CrtQ2 y CrtO2, fue de máximo cuatro (**Figura 10**); es decir, representó menos del dos por ciento de las proteínas encontradas (**Figura 12**). La proteína CrtM participa en la formación de carotenoides de 30 átomos de carbono a partir de dos moléculas de farnesil difosfato (C_{15}) (Klassen, 2010). De acuerdo con Pelz (*et al.,* 2006), la proteína CrtM forma parte del operón CrtO2P2Q2MN, el cual favorece la producción del factor de virulencia estafiloxantina en *Staphylococcus aureus* (Liu et al., 2005; Pelz et al., 2005). En este estudio se observó que todos los representantes de *S. aureus* presentaron la misma abundancia de las proteínas del operón CrtO2P2Q2MN, excepto Staphylococcus aureus MW2 que fue la única que no presentó a la proteína CrtN. La abundancia tan reducida de la proteínas que conforman el operón pudo estar influenciada con una representación mínima de bacterias en las que comúnmente se encuentra la proteína o con una conservación limitada que se limita a géneros específicos (ej. *S. aureus*) (Pelz et al., 2005). En trabajos previos (Klassen, 2010), la formación de carotenoides de 30 átomos de carbono fue analizada sólo con la proteína CrtM; sin embargo, en el presente estudio también se consideraron las proteínas CrtN, CrtP2, CrtQ2, y CrtO2. Esto se realizó con la finalidad de evaluar la presencia del operón CrtO2P2Q2MN en diversos bacterias, incluyendo *S. aureus* (Pelz et al., 2005).

La proteína CrtM se encontró distribuida en los phyla Firmicutes (*Staphylococcus aureus)* y Actinobacteria (*Streptomyces coelicolor)* (**Figura 16)**. La presencia de la proteína CrtM en los géneros de *Staphylococcus* y *Streptomyces*, confirma los encontrado por Umeno (*et al.*, 2002). La proteína CrtN fue encontrada sólo en un miembro del phylum Firmicutes (*Staphylococcus aureus,* excepto en *S. aureus* MW2) (**Figura 16**). La presencia de la proteína CrtN ha sido descrita en *S. aureus* para la formación del factor de virulencia estafiloxantina (Wieland et al., 1994). Las proteínas CrtP2, CrtQ2 y CrtO2, se encontraron en los mismos representantes del phylum Firmicutes (*Staphylococcus aureus)*

(**Figura 16**).

Síntesis de fitoeno

Fitoeno sintasa (CrtB), [EC:2.5.1.32]: La abundancia de homólogos de la proteína CrtB fue de nueve (**Figura 10**), es decir, fue encontrada en menos del 3% de las proteínas totales (**Figura 12**). Las proteínas fitoeno sintasa tienen homología con las diapofitoeno sintasa y escualeno sintasa, compartiendo regiones consenso entre sí (Sieiro et al., 2003)

La proteína CrtB fue encontrada en los phyla de Bacteroidetes (*Bacteroides sp., Porphyromonas sp.*), Chlorobi (*Chlorobium tepidum*), Deinococcoci (*Deinococcus radiodurans*), Cyanobacteria (*Gloeobacter sp., Synechococcus., Synechocystis sp., Nostoc sp.,*) y en las clases Alphaproetobacteria (*Rhodopseudomonas palustris*), Betaproteobacteria (*Chromobacterium violaceum*) y Gammaproteobacteria (*Photorhabdus luminescens*). En diversos estudios se ha reportado la presencia de la proteína CrtB en los phyla Cyanobacteria (*Synechococcus*)(Armstrong, 1997). Aunque la proteína CrtB ha sido encontrada en representantes de los géneros *Agrobacterium, Streptomyces* y *Thermus*, en este estudio no se encontraron proteínas relacionadas dentro del rango de selección del perfil HMM de CrtB (Hoshino et al., 1993). Se encontró que los *bit score* de algunos homólogos estaban fuera del rango de selección y no pudieron ser consideradas aunque la bibliografía las mencionara (*Agrobacterium fabrum* C58, *bit score* 63.1), *Streptomyces coelicolor* A3 2 (*bit score* 86.6 a 221.9), *Streptomyces avermitilis* MA 4680 (*bit score* 66.6 a 222) y *Thermus thermophilus* HB27 (*bit score* 237.5). La proteína CrtB, puede encontrarse en operones junto a las proteínas CrtA, CrtC, CrtD, CrtE, CrtF y CrtI y en algunos casos, en operones que participan en la biosíntesis de bacterioclorofila, como en *Rhodobacter capsulatus* (Gallagher, Matthews, Li, & Wurtzel, 2004). La copresencia de CrtB, junto con las proteínas antes mencionadas, permite la producción del compuesto esferoideno. Y junto

con CrtE, CrtI, CrtX, CrtY y CrtZ, permiten la síntesis de zeaxantina y criptoxantina (Armstrong, 1997). Se propone que para posteriores estudios, se considere a otras bacterias como *Erwina herbicola, Erwina uredevora, Flavobacterium sp., Myxococcus xanthus, Rhodobacter capsulatus, R. sphaeroides, Streptomyces griseus,* para evaluar la presencia de CrtB (Armstrong, 1997; Lang, Cogdell, Gardiner, & Hunter, 1994; Sieiro et al., 2003).

Proteína bifuncional CrtB2 fitoeno sintasa/isoprenil transferasa [EC:2.5.1.32] / [EC= 2.5.1.-]: La proteína CrtB2 fue encontrada en el phylum de las actinobacterias (*Bifidobacterium sp., Mycobacterium sp., Corynebacterium* y *Streptomyces avermitilis*) y en la clase de las alfaproteobacterias (*Rhodopseudomonas palustris* y *Bradyrhizobium japonicum).* Aunque no se tiene registro previo sobre la presencia de la proteína CrtB2 en los grupos antes mencionados, se propone que podría ser determinante en algunos estilos de vida para un mayor producción de carotenoides. Se considera que la proteína CrtB2 es importante para las bacterias que la producen debido a que es capaz de asociarse con otras proteínas y puede aumentar la síntesis de los carotenoides relacionados, tal como ocurre con la proteína bifuncional CrtBY presente en hongos (Dogbo, Laferrière, Harlingue, & Camara, 1988; Sieiro et al., 2003). Sin embargo, en bacterias, la proteína CrtB2 no ha sido caracterizada más allá de predicciones por SwissProt en el proteoma de *Streptomyces coelicolor* A3(2) (NC_003903.1). En este estudio, las proteínas CrtB y CrtB2, se usaron como representantes de la síntesis de fitoeno, aunque comúnmente es usada sólo la proteína CrtB (Phadwal, 2005). Respecto a otros estudios (Phadwal 2005), en los que se representó la formación de carotenoides con 40 átomos de carbono con la formación de GGPP, en el presente estudio también se consideró la presencia de CrtE.

Deshidrogenación de fitoeno

Fitoeno desaturasa, formadora de neurosporeno (CrtI), [EC=1.3.99.29]: La proteína CrtI fue encontrada en las clases Alphaproteobacteria (*Rhodopseudomonas palustris*) y Gammaproteobacteria (*Photorhabdus luminescens*) del phylum Proteobacteria. En diversos estudios, la proteína CrtI ha sido reportada en los phyla Cyanobacteria y para la Alphaproteobacteria (*Rhodobacter sphaeroides*) (Lang, Cogdell, Gardiner, & Hunter, 1994b). En este estudio, se ha confirmado la presencia de CrtI en las Alphaproteobacterias. **Como en los casos anteriores, deberías intentar dar una explicación al por qué de las presencias, y no sólo quedarte con mencionarlas. Creo que deberías revisar más la bioquímica de los organismos**

15-cis-fitoeno desaturasa (CrtP), [EC=1.3.5.5]: La proteína CrtP fue encontrada sólo en el phylum Cyanobacteria *(Synechococcus sp., Synechocystis sp., Nostoc sp.,* y *Prochlorococcus marinus)*. La presencia de la proteína CrtP en cianobacterias favorece la formación de licopeno a partir de fitoeno mediante cuatro reacciones consecutivas de desaturación; sin embargo, en algunos casos las cuatro desaturaciones pueden ser catalizadas por sólo una enzima (CrtI), tal es el caso de la cianobacteria *Gloebacter violaceus* (Mulkidjanian et al., 2006; Tsuchiya et al., 2005). Lo anterior, podría ser una de las razones por la cuales no todas las proteínas presentan la proteína CrtP en una reacción tan importante como es la desaturación del fitoeno. En la literatura, la presencia de de CrtP ha sido confirmada en *Synechococcus elongatus* (C. Schneider et al., 1997).

All-trans-zeta-caroteno desaturasa (CrtIb), [EC=1.3.99.26]: Son un grupo de enzimas que catalizan los pasos de desaturación desde fitoeno a licopeno. La proteína CarC, ha sido reportada en estudios previos con la Deltaproteobacteria, *Myxococcus xanthus* (Iniesta et al., 2007), sin embargo,

para este estudio no encontramos homólogos de CarC en el grupo de las Deltaproteobacteria, pero se encontraron secuencias de CarC en el grupo Cyanobacteria, Thermotoga, Acidobacteria, Firmicutes y en la mayoría de los órdenes del Proteobacteria.

Fitoeno desaturasa formadora de zeta-caroteno (CrtIa), [EC=1.3.99.29]: La proteína CarA2 o CrtIa, ha sido caracterizada en la Deltaproteobacteria, *Myxococcus xanthus* (Iniesta et al., 2007). En este trabajo, se encontraron secuencias homólogas de CarA2 en Cianobacteria, Acidobacteria y Gammaproteobacteria.

Ciclización de licopeno

Licopeno beta ciclasa (CrtL), [EC:5.5.1.19]: La presencia de la enzima licopeno beta ciclasa (CrtL) ha sido confirmada en la Cianobaceteria del género Synechococcus (Francis X Cunningham et al., 1994). Adicionalmente, se observa su presencia en Actinobacteria, Deinococcus-Thermus y en Gammaproteobacteria.

Licopeno ciclasa tipo CruP, [EC: 5.5.1.19]: Las enzimas licopeno ciclasa tipo CruA y CruP; y la licopeno epsilon-ciclasa (LcyE), han sido descritas en *Chlorobium tepidium* (Young, Bauer, Williams, & Marrs, 1989), sin embargo, en este trabajo sólo se encontró a CruP en el phylum Cyanobacteria (*Nostoc sp, Synechocystis*).

Actividades enzimáticas posteriores a la ciclización de licopeno

La **Isoreniereteno sintasa (CrtU) [EC:1.-.-.-]** se encontró distribuida en la especie *Streptomyces*

avermitilis, lo cual confirma lo descrito para esta enzima en el género *Streptomyces*, particularmente en *Streptomyces griseus* (Krügel, Krubasik, Weber, Peter-Saluz, & Sandmann, 1999). Sin embargo, *S. coelicolor* no presentó a la proteína CrtU, lo cual habla sobre la variación de la distribución de una sola proteína en un mismo género.

Beta-caroteno cetolasa tipo CrtW (CrtW): Se ha reportado en la Cianobacteria *Anabaena* sp. PCC 7120, CrtW tiene actividad de beta-caroteno cetolasa tipo CrtW, que permiten la formación de cetomixol (Liang et al., 2006; Mochimaru, Masukawa, & Takaichi, 2005). Esto reafirma lo encontrado en el presente estudio, sobre su amplia presencia en Cianobacteria. También se encontraron homólogos CrtW en Actinobacteria y en Betaproteobacteria.

Clorobacteno glucosiltransferasa (CruC): La presencia de la proteína ha sido confirmada en *Chlorobium tepidium* (Maresca & Bryant, 2006). En ningún otro phylum que no fuera Chlorobi, se encontró esta transferasa. Maresca (*et al.*, 2006), propone que CruC y CruD, se encuentran presentes en todas las bacterias verde-azules sulfurosas y fototróficas anoxigénicas filamentosas, sin embargo, en este trabajo sólo se encontraron en las especies ya mencionadas.

Caroteno 1,2-hidratasa (CrtC): La proteína ha sido encontrada en *Rubrivivax gelatinosus* (Betaproteobacteria), sin embargo, no se encontró en ningún otro representante de las Betaproteobacteria (Hiseni et al., 2011).

Formación de carotenoides acíclicos

Dimetilesferoideno O-metiltransferasa (CrtF): La proteína ha sido descrita en

Alphaproteobacteria, *Rhodobacter capsulatus* o *Rhodopseudomonas capsulatus* (Scolnik et al., 1980; Young et al., 1989). En este estudio, se ha encontrado la proteína CrtF en uno de los representantes *Rhodopseudomonas palustris* (Alfaproteobacteria). En otros estudios, la proteína CrtF se ha encontrado en Planctomicetes, Cianobacteria y Acidobacterium.

1-Hidroxicaroteno 3,4-desaturasa (CrtD): La proteína ha sido reportada en la Betaproteobacteria *Rubrivivax gelatinous* (Steiger, Astier, & Sandmann, 2000), sin embargo, en este estudio no se encontró ninguna secuencia homóloga en el orden Betaproteobacteria

Esferoideno monoooxigenasa (CrtA) [EC:1.14.15.9]: La proteína CrtA, ha sido descrita en las Alphaproteobacteria, *Rhodobacter capsulatus, R. sphaeroides*, así como en la Betaproteobacteria *Rubrivivax gelatinosus* (Armstrong et al., 1989; Gerjets et al., 2009). Sin embargo, en este estudio no se localizó ninguna proteína entre los géneros representativos de las Alfaproteobacterias y Betaproteobacterias analizadas. Esto podría deberse a que la proteína CrtA tiene una distribución filogenética muy restringida y que ninguna de las especies usadas representa su verdadera distribución. Sin embargo, se encontró en la Actinobacteria *Streptomyces coelicolor.*

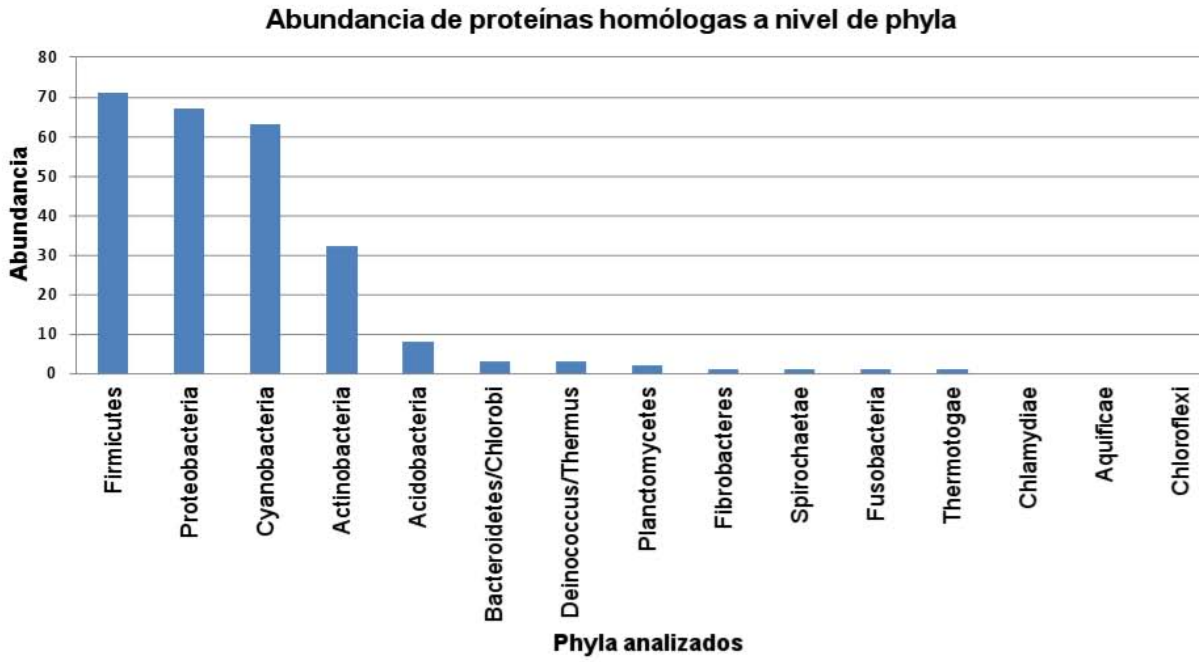
Abundancia de las proteínas de la BC a nivel de phylum

Los phyla con mayor abundancia de proteínas de la BC fueron Firmicutes (71), Cyanobacteria (63) y el phylum Proteobacteria, particularmente la clase de Gammaproteoacteria (37). Como se consideró que los phyla con mayor abundancia podrían estar sobre representados por la cantidad de organismos analizados, también se reportó la abundancia normalizada (**Tabla 2)**.

La abundancia normalizada es el resultado de dividir la abundancia de las proteínas de la BC entre el número de proteomas analizados. El phylum de cianobacterias fue el mayor en abundancia relativa (7.85), lo que podría estar relacionado con la capacidad de los carotenoides para actuar como pigmentos accesorios y prevenir el daño fotooxidativo durante la fotosíntesis (Liang et al., 2006; Mulkidjanian et al., 2006; Tóth et al., 2015). El valor de abundancia normalizada del phylum Cyanobacteria fue seguido de los phyla de Acidobacteria (4) y Actinobacteria (2.28) (**Tabla 2**). Una representación gráfica de la **Tabla 2**, se muestra en la **Figura 13**. Los phyla que no presentaron ninguna proteína de la BC fueron Chlamydiae, Aquificae, Chloroflexi y Campylobacterales. Chlamydiae, esto podría deberse a que en sus estilos de vida no son requeridos los carotenoides o se han perdido a lo largo de la divergencia evolutiva. También podría ocurrir que los perfiles HMM no hallan sido capaces de reconocer homólogos remotos de la BC.

Tabla 2. Abundancia y abundancia relativa de las proteínas de la biosíntesis de carotenoides en los phyla analizados. La abundancia normalizada representa el cociente entre la abundancia de las proteínas de la BC y la cantidad de proteomas analizados. En el caso de las proteobacterias, las abundancias se desglosan en sus clases. Los phyla se encuentran organizados por su valor de abundancia normalizada, de mayor a menor.

Phylum -Clase	Abundancia de las proteínas de la BC	Número de proteomas analizados	Abundancia normalizada *
Cyanobacteria	63	8	7.87
Acidobacteria	8	2	4
Actinobacteria	32	14	2.28
Planctomycetes	2	1	2
Firmicutes	71	39	1.80
Deinococcus/Thermus	3	2	1.5
Proteobacteria	67	63	1.06
-Alphaproteobacteria	16	11	1.45
-Betaproteobacteria	11	8	1.37
-Gammaproteobacteria	37	36	1.02
-Deltaproteobacteria	3	3	1
-Epsilonproteobacteria (Campylobacterales)	0	5	0
Thermotogae	1	1	1
Fusobacteria	1	1	1
Bacteroidetes/Chlorobi	3	3	1
Fibrobacteres	1	1	1
Spirochaetae	1	5	0.2
Chloroflexi	0	1	0
Aquificae	0	1	0
Chlamydiae	0	7	0
Total	253	149	-



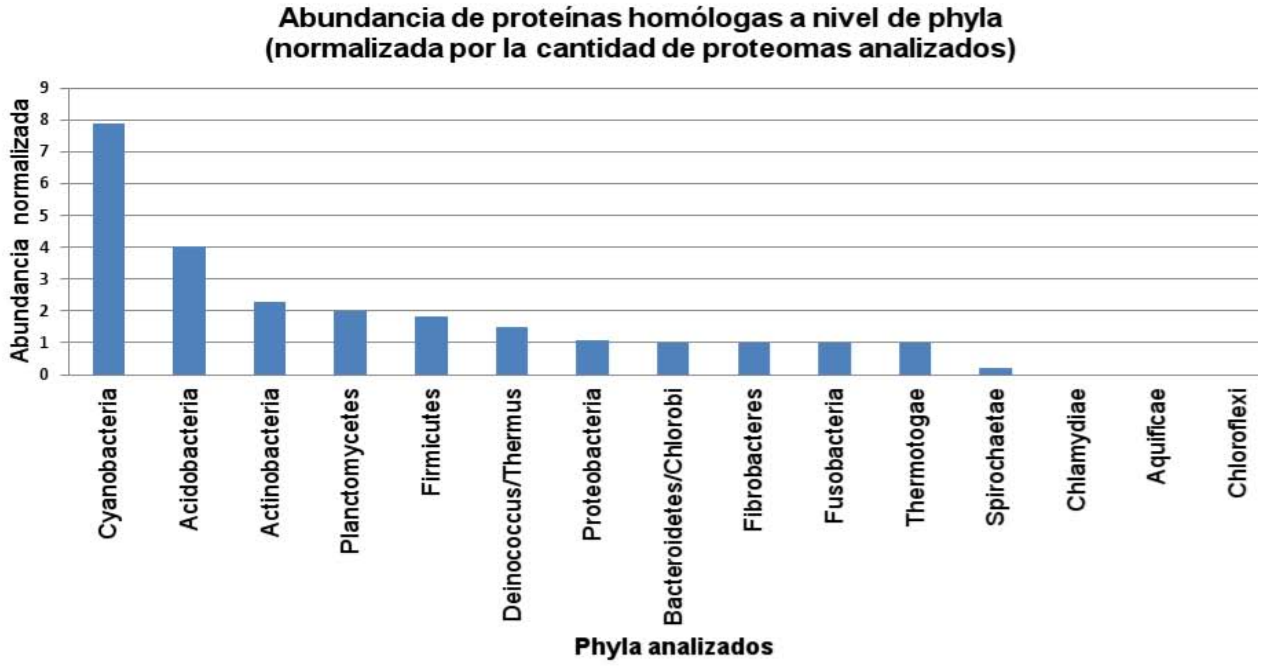


Figura 13 y 14. Abundancia y abundancia normalizada de las proteínas homólogas de la biosíntesis de carotenoides, respectivamente. La abundancia de los phyla depende de la cantidad de bacterias analizadas.

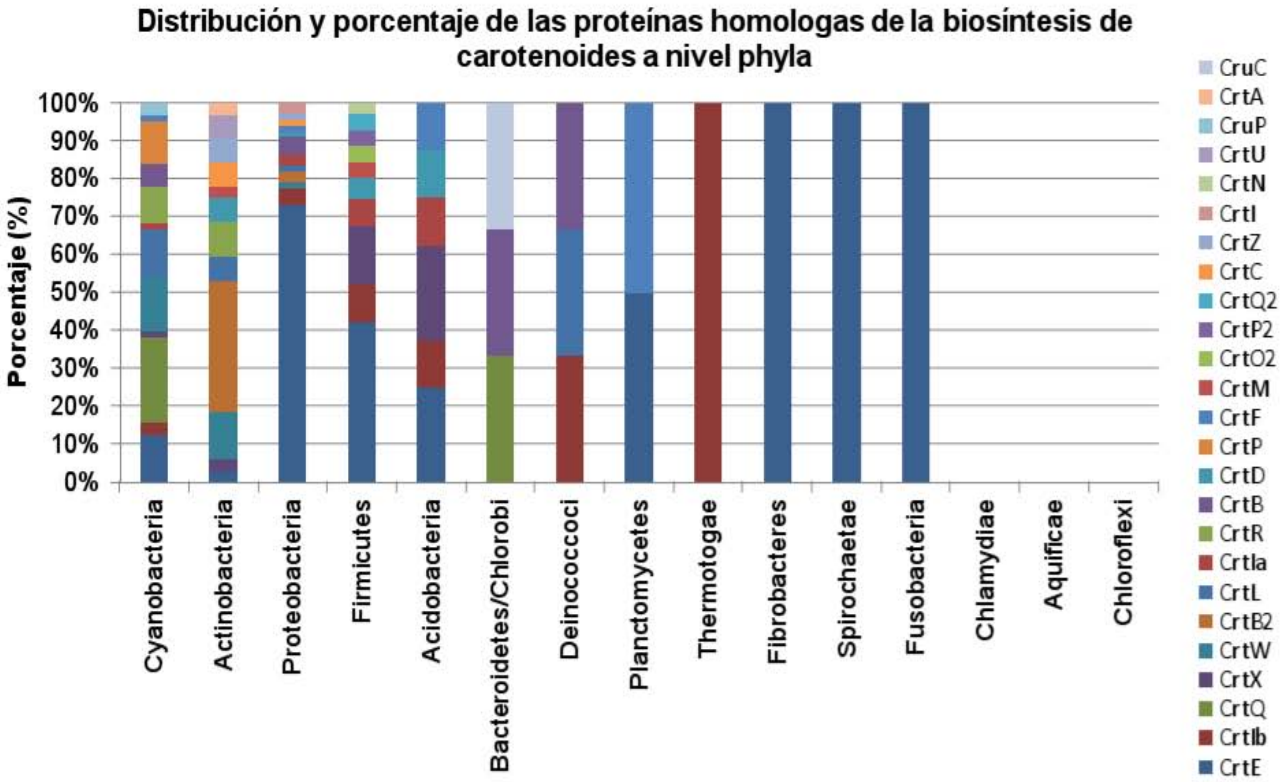


Figura 15. Distribución y proporción de las proteínas de la BC en los principales phyla bacterianos. Los phyla se muestran ordenados de mayor a menor diversidad. La abundancia de las proteínas es representada mediante porcentajes y los porcentajes son independientes entre los phyla. Se puede observar que la distribución y abundancia de las proteínas de la BC es diferente en los phyla estudiados.

Abundancia de las proteínas de la BC, a nivel de proteoma

La abundancia relativa de las proteínas de la BC, respecto al total de proteínas presentes en el proteoma, se muestra en la **Figura 15**. La abundancia de las proteínas de la BC en cada uno de los

proteomas analizados se representa en la **Figura 16**. En el **Anexo 4** se muestra la matriz de datos con la abundancia relativa de las proteínas de la BC, en cada proteoma analizado. La cianobacteria *Nostoc* sp. PCC 7120, fue la bacteria con mayor abundancia de proteínas de la BC (15). Esto podría deberse a que las dicha cianobacteria conserva gran parte de la ruta de la BC para facilitar la diversificación de compuestos carotenoides y con esto ayudar a la captación de luz y protección del daño fotooxidativo (Liang et al., 2006). Sin embargo, la abundancia relativa de las proteínas de la BC, resulto ser menor al 1 % del total de proteínas presentes en su proteoma (6,129) (**Figura 16**). Esto podría deberse a que la abundancia de las proteínas es suficiente para la funcionalidad de la BC, aunque podrían existir la probabilidad de que una sobrerepresentación de genes codificantes de proteínas de la BC sea necesaria para la acumulación de carotenoides. Lo anterior podría tener repercusiones en la adaptabilidad de las bacterias de distintos hábitos (ej. extremófilas, fotosintéticas).

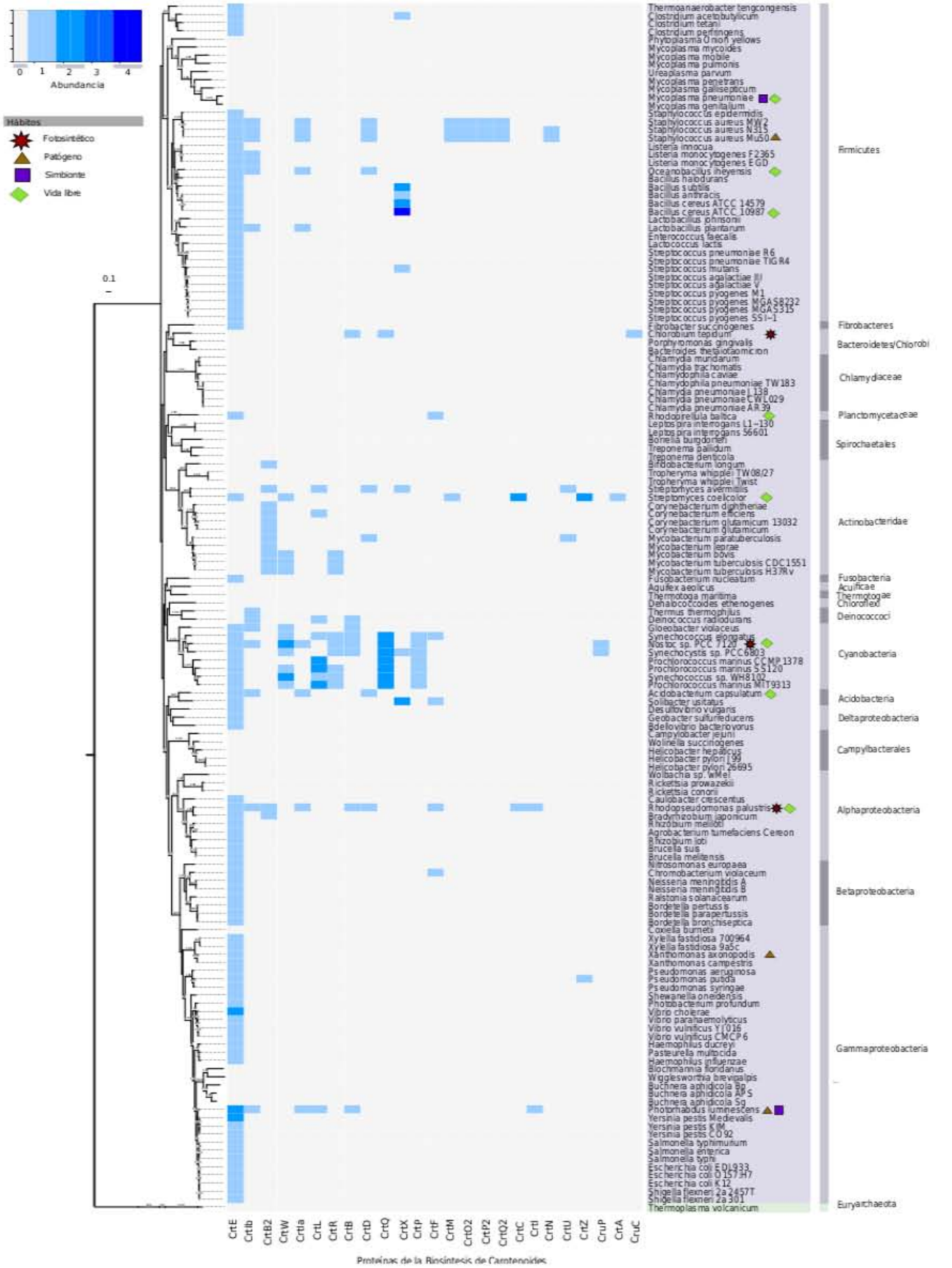

0
1
2
3
4
Abundancia
Hábitos
Fotosintético
Patógeno
Simbionte
Vida libre
0.1
Thermoanaerobacter tengcongensis
Clostridium acetobutylicum
Clostridium tetani
Clostridium perfringens
Phytoplasma Onion yellows
Mycoplasma mycoides
Mycoplasma mobile
Mycoplasma pulmonis
Ureaplasma parvum
Mycoplasma penetrans
Mycoplasma gallisepticum
Mycoplasma pneumoniae
Mycoplasma genitalium
Staphylococcus epidermidis
Staphylococcus aureus MW2
Staphylococcus aureus N315
Staphylococcus aureus Mu50
Listeria innocua
Listeria monocytogenes F2365
Listeria monocytogenes EGD
Oceanobacillus iheyensis
Bacillus halodurans
Bacillus subtilis
Bacillus anthracis
Bacillus cereus ATCC 14579
Bacillus cereus ATCC 10987
Lactobacillus johnsonii
Lactobacillus plantarum
Enterococcus faecalis
Lactococcus lactis
Streptococcus pneumoniae R6
Streptococcus pneumoniae TIGR4
Streptococcus mutans
Streptococcus agalactiae III
Streptococcus agalactiae V
Streptococcus pyogenes M1
Streptococcus pyogenes MGAS8232
Streptococcus pyogenes MGAS315
Streptococcus pyogenes SS1−1
Fibrobacter succinogenes
Chlorobium tepidum
Porphyromonas gingivalis
Bacteroides thetaiotaomicron
Chlamydia muridarum
Chlamydia trachomatis
Chlamydophila caviae
Chlamydophila pneumoniae TW183
Chlamydia pneumoniae LJ138
Chlamydia pneumoniae CWL029
Chlamydia pneumoniae AR39
Rhodopirellula baltica
Leptospira interrogans L1−130
Leptospira interrogans 56601
Borrelia burgdorferi
Treponema pallidum
Treponema denticola
Bifidobacterium longum
Tropheryma whipplei TW08/27
Tropheryma whipplei Twist
Streptomyces avermitilis
Streptomyces coelicolor
Corynebacterium diphtheriae
Corynebacterium efficiens
Corynebacterium glutamicum 13032
Corynebacterium glutamicum
Mycobacterium paratuberculosis
Mycobacterium leprae
Mycobacterium bovis
Mycobacterium tuberculosis CDC1551
Mycobacterium tuberculosis H37Rv
Fusobacterium nucleatum
Aquifex aeolicus
Thermotoga maritima
Dehalococcoides ethenogenes
Thermus thermophilus
Deinococcus radiodurans
Gloeobacter violaceus
Synechococcus elongatus
Nostoc sp. PCC 7120
Synechocystis sp. PCC6803
Prochlorococcus marinus CCMP1378
Prochlorococcus marinus SS120
Synechococcus sp. WH8102
Prochlorococcus marinus MIT9313
Acidobacterium capsulatum
Solibacter usitatus
Desulfovibrio vulgaris
Geobacter sulfurreducens
Bdellovibrio bacteriovorus
Campylobacter jejuni
Wolinella succinogenes
Helicobacter hepaticus
Helicobacter pylori J99
Helicobacter pylori 26695
Wolbachia sp. wMel
Rickettsia prowazekii
Rickettsia conorii
Caulobacter crescentus
Rhodopseudomonas palustris
Bradyrhizobium japonicum
Rhizobium meliloti
Agrobacterium tumefaciens Cereon
Rhizobium loti
Brucella suis
Brucella melitensis
Nitrosomonas europaea
Chromobacterium violaceum
Neisseria meningitidis A
Neisseria meningitidis B
Ralstonia solanacearum
Bordetella pertussis
Bordetella parapertussis
Bordetella bronchiseptica
Coxiella burnetii
Xylella fastidiosa 700964
Xylella fastidiosa 9a5c
Xanthomonas axonopodis
Xanthomonas campestris
Pseudomonas aeruginosa
Pseudomonas putida
Pseudomonas syringae
Shewanella oneidensis
Photobacterium profundum
Vibrio cholerae
Vibrio parahaemolyticus
Vibrio vulnificus YJ016
Vibrio vulnificus CMCP6
Haemophilus ducreyi
Pasteurella multocida
Haemophilus influenzae
Blochmannia floridanus
Wigglesworthia brevipalpis
Buchnera aphidicola Bp
Buchnera aphidicola APS
Buchnera aphidicola Sg
Photorhabdus luminescens
Yersinia pestis Medievalis
Yersinia pestis KIM
Yersinia pestis CO92
Salmonella typhimurium
Salmonella enterica
Salmonella typhi
Escherichia coli EDL933
Escherichia coli O157H7
Escherichia coli K12
Shigella flexneri 2a 2457T
Shigella flexneri 2a 301
Thermoplasma volcanicum
Firmicutes
Fibrobacteres
Bacteroidetes/Chlorobi
Chlamydiaceae
Planctomycetaceae
Spirochaetales
Actinobacteridae
Fusobacteria
Acuificae
Thermotogae
Chloroflexi
Deinococci
Cyanobacteria
Acidobacteria
Deltaproteobacteria
Campylbacterales
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Euryarchaeota
CrtE
CrtIb
CrtB2
CrtW
CrtIa
CrtL
CrtR
CrtB
CrtD
CrtQ
CrtX
CrtP
CrtF
CrtM
CrtO2
CrtP2
CrtO2
CrtC
CrtI
CrtN
CrtU
CrtZ
CruP
CrtA
CruC
Proteínas de la Biosíntesis de Carotenoides

Figura 16. Distribución de las proteínas de la BC en los principales phyla de bacterias. En la parte derecha de la figura se muestran los proteomas analizados, agrupados en los respectivos phyla. En la parte central de la figura, se representa la presencia y ausencia de secuencias homólogas de la BC en celdas de color azul. A mayor abundancia, el color se muestra más obscuro, lo que representa un máximo de 4 copias. En la parte inferior de la figura se muestran las siglas que denotan a las proteínas de la BC analizadas, las cuales están organizadas por columnas. Las columnas se ordenaron por su abundancia, de izquierda a derecha. En la parte izquierda de la figura, se muestra la filogenia del árbol de la vida propuesta por Ciccarelli (*et al*., 2006). En la parte superior izquierda, se muestra la escala de colores para la cantidad de copias encontradas en cada proteoma. En seguida, se muestra un cuadro en el que se muestran las principales categorías de hábitos de vida (*ej.* fotosintético, patógeno, simbionte, vida libre) asociados a las bacterias analizadas.

Abundancia de proteínas de la BC con genes de copia única

Se encontró que la mayoría de las proteínas analizadas (207 / 253) corresponden a proteínas codificadas por genes de copia única en el proteoma. Esto es interesante, respecto a la presencia de proteínas codificadas por genes parálogos, en las que la presencia de varias copias en el genoma puede estar relacionada con la utilidad de incrementar la producción de dicho compuesto (Gottlieb, Albermann, & Sprenger, 2014). En plantas, se han encontrado genes parálogos (PSY, PSY1, PSY2) de la BC, que podrían estar implicados en la acumulación de carotenoides en el endospermo (Finn et al., 2011). En bacterias no se han descrito casos similares, sin embargo podría ocurrir lo mismo si se cuenta con más de una copia por gen tal como ocurre en plantas. Asimismo, la presencia de varias copias de

un mismo gen, aparte de generar potenciales beneficios como aumentar la producción de un determinado compuesto, también es una fuente potencial de neofuncionalización (Assis & Bachtrog, 2013). En oposición a lo anterior, Adler y colaboradores (2014), sugieren que tener más de una copia de un gen, no siempre representa un potencial beneficio, sino que puede representar costos metabólicos adicionales; principalmente por incrementar la regulación de las interacciones moleculares en la ruta (Adler, Anjum, Berg, Andersson, & Sandegren, 2014). En el caso de las bacterias, en las que se encontraron copias múltiples, se consideró que podrían estar codificadas en plásmidos. Sin embargo, no siempre se presentaron las copias de las proteína en los plásmidos. Tal es el caso de la bacteria *Bacillus cereus* ATCC 10987, en la que se encontraron cuatro secuencias para la misma proteína (CrtX) en el genoma cromosomal. La presencia de proteínas de BC en los plásmidos de bacterias, podría estar relacionada con la adecuación de la bacteria por estrés en el ambiente debido a que en los plásmido pueden alojarse genes codificantes que favorecen la adaptabilidad de bacterias (Khaneja et al., 2010a).

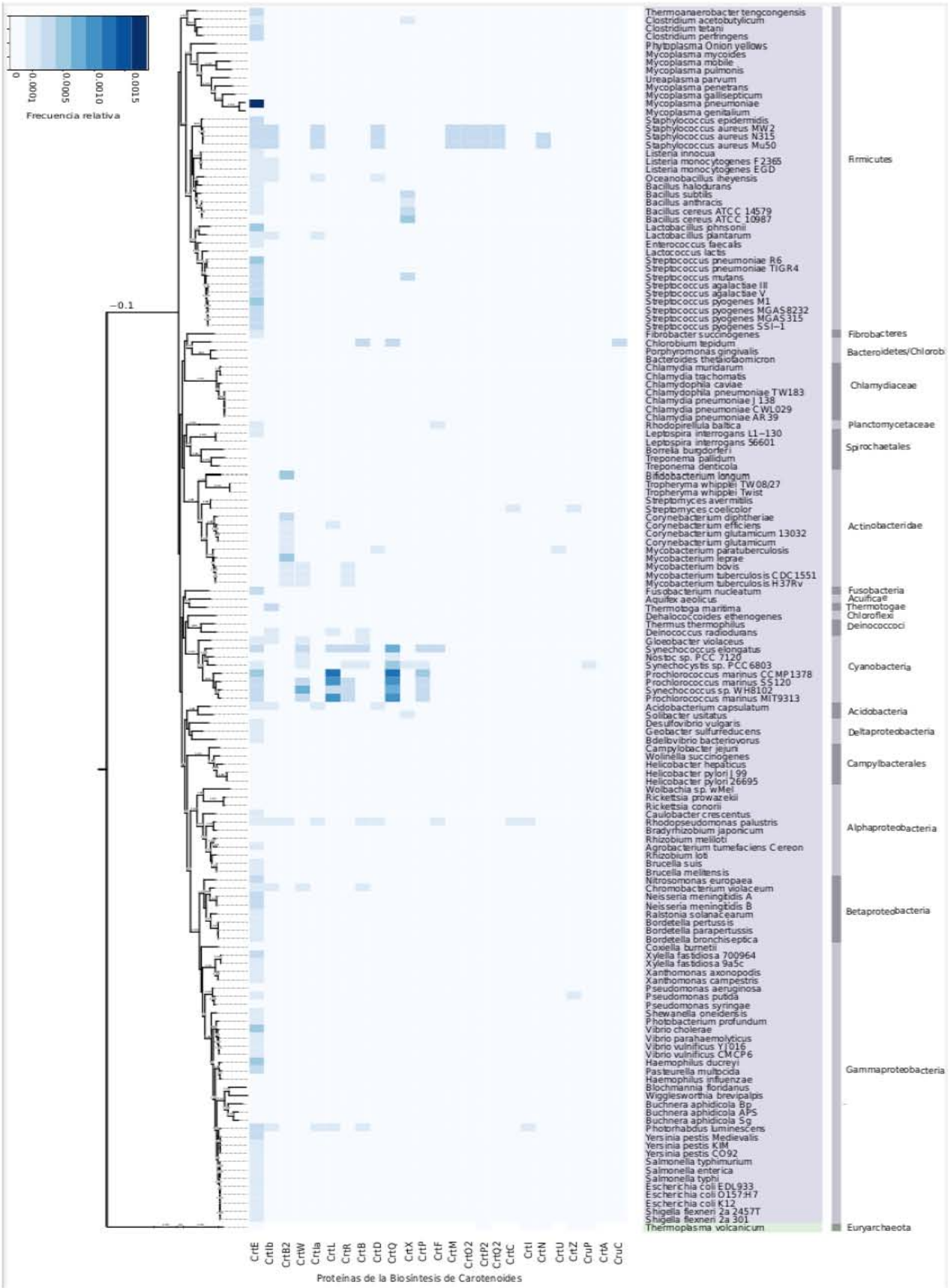

0
0.0001
0.0005
0.0010
0.0015
Frecuencia relativa
0.1
Thermoanaerobacter tengcongensis
Clostridium acetobutylicum
Clostridium tetani
Clostridium perfringens
Phytoplasma Onion yellows
Mycoplasma mycoides
Mycoplasma mobile
Mycoplasma pulmonis
Ureaplasma parvum
Mycoplasma penetrans
Mycoplasma gallisepticum
Mycoplasma pneumoniae
Mycoplasma genitalium
Staphylococcus epidermidis
Staphylococcus aureus MW2
Staphylococcus aureus N315
Staphylococcus aureus Mu50
Listeria innocua
Listeria monocytogenes F2365
Listeria monocytogenes EGD
Oceanobacillus iheyensis
Bacillus halodurans
Bacillus subtilis
Bacillus anthracis
Bacillus cereus ATCC 14579
Bacillus cereus ATCC 10987
Lactobacillus johnsonii
Lactobacillus plantarum
Enterococcus faecalis
Lactococcus lactis
Streptococcus pneumoniae R6
Streptococcus pneumoniae TIGR4
Streptococcus mutans
Streptococcus agalactiae III
Streptococcus agalactiae V
Streptococcus pyogenes M1
Streptococcus pyogenes MGAS8232
Streptococcus pyogenes MGAS315
Streptococcus pyogenes SSI−1
Fibrobacter succinogenes
Chlorobium tepidum
Porphyromonas gingivalis
Bacteroides thetaiotaomicron
Chlamydia muridarum
Chlamydia trachomatis
Chlamydophila caviae
Chlamydophila pneumoniae TW183
Chlamydia pneumoniae J138
Chlamydia pneumoniae CWL029
Chlamydia pneumoniae AR39
Rhodopirellula baltica
Leptospira interrogans L1−130
Leptospira interrogans 56601
Borrelia burgdorferi
Treponema pallidum
Treponema denticola
Bifidobacterium longum
Tropheryma whipplei TW08/27
Tropheryma whipplei Twist
Streptomyces avermitilis
Streptomyces coelicolor
Corynebacterium diphtheriae
Corynebacterium efficiens
Corynebacterium glutamicum 13032
Corynebacterium glutamicum
Mycobacterium paratuberculosis
Mycobacterium leprae
Mycobacterium bovis
Mycobacterium tuberculosis CDC1551
Mycobacterium tuberculosis H37Rv
Fusobacterium nucleatum
Aquifex aeolicus
Thermotoga maritima
Dehalococcoides ethenogenes
Thermus thermophilus
Deinococcus radiodurans
Gloeobacter violaceus
Synechococcus elongatus
Nostoc sp. PCC 7120
Synechocystis sp. PCC 6803
Prochlorococcus marinus CCMP1378
Prochlorococcus marinus SS120
Synechococcus sp. WH8102
Prochlorococcus marinus MIT9313
Acidobacterium capsulatum
Solibacter usitatus
Desulfovibrio vulgaris
Geobacter sulfurreducens
Bdellovibrio bacteriovorus
Campylobacter jejuni
Wolinella succinogenes
Helicobacter hepaticus
Helicobacter pylori J99
Helicobacter pylori 26695
Wolbachia sp. wMel
Rickettsia prowazekii
Rickettsia conorii
Caulobacter crescentus
Rhodopseudomonas palustris
Bradyrhizobium japonicum
Rhizobium meliloti
Agrobacterium tumefaciens Cereon
Rhizobium loti
Brucella suis
Brucella melitensis
Nitrosomonas europaea
Chromobacterium violaceum
Neisseria meningitidis A
Neisseria meningitidis B
Ralstonia solanacearum
Bordetella pertussis
Bordetella parapertussis
Bordetella bronchiseptica
Coxiella burnetii
Xylella fastidiosa 700964
Xylella fastidiosa 9a5c
Xanthomonas axonopodis
Xanthomonas campestris
Pseudomonas aeruginosa
Pseudomonas putida
Pseudomonas syringae
Shewanella oneidensis
Photobacterium profundum
Vibrio cholerae
Vibrio parahaemolyticus
Vibrio vulnificus YJ016
Vibrio vulnificus CMCP6
Haemophilus ducreyi
Pasteurella multocida
Haemophilus influenzae
Blochmannia floridanus
Wigglesworthia brevipalpis
Buchnera aphidicola Bp
Buchnera aphidicola APS
Buchnera aphidicola Sg
Photorhabdus luminescens
Yersinia pestis Medievalis
Yersinia pestis KIM
Yersinia pestis CO92
Salmonella typhimurium
Salmonella enterica
Salmonella typhi
Escherichia coli EDL933
Escherichia coli O157:H7
Escherichia coli K12
Shigella flexneri 2a 2457T
Shigella flexneri 2a 301
Thermoplasma volcanicum
Firmicutes
Fibrobacteres
Bacteroidetes/Chlorobi
Chlamydiaceae
Planctomycetaceae
Spirochaetales
Actinobacteridae
Fusobacteria
Aquificae
Thermotogae
Chloroflexi
Deinococci
Cyanobacteria
Acidobacteria
Deltaproteobacteria
Campylbacterales
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Euryarchaeota
CrtE
CrtIb
CrtB2
CrtW
CrtIa
CrtL
CrtR
CrtB
CrtD
CrtQ
CrtX
CrtP
CrtF
CrtM
CrtO2
CrtP2
CrtO2
CrtC
CrtI
CrtN
CrtU
CrtZ
CruP
CrtA
CruC
Proteinas de la Biosíntesis de Carotenoides

Figura 17. Abundancia relativa de las proteínas de la BC, respecto al total de proteínas presentes en los proteomas de bacterias analizados. En la parte derecha, se muestran los proteomas analizados, agrupados por phyla. En la parte central se muestra la abundancia relativa, en escala de color azul. En la parte izquierda de la figura se muestra la filogenia del árbol de la vida (itol) generada por Ciccarelli. En la parte inferior se muestran las siglas que hacen referencia a las proteínas de la BC analizadas.

Implicaciones de los carotenoides en la biología y ecología de las bacterias

La distribución de carotenoides no se había estudiado en representantes no fotosintéticos de bacterias como Firmicutes, Fibrobacteres, Chlamydiae, Planctomycetace, Spirochaetales, Actinobacteria, Fusobacteria, Acuificae, Deinococcoci, Acidobacteria, Campylobacterales y Proteobacteria. Respecto a la relación de biosíntesis de carotenoides con los hábitos y formas de vida que desempeñan las bacterias en su ambiente, se encontró una gran diversidad de actividades relacionadas con los carotenoides (**Figura 16**). En Cianobacteria la presencia de proteínas de la BC está relacionada con la fotosíntesis y con propiedades antioxidantes (Hashtroudi, Shariatmadari, Riahi, & Ghassempour, 2013; Mulkidjanian & Galperin, 2013). Al igual que las cianobacterias, las bacterias fotoautotróficas del phylum Chlorobi utilizan a los carotenoides como fuente de protección y para realizar la fotosíntesis, aunque en algunos casos se puede prescindir de éstos sin afectar al proceso de captación de luz (Frigaard, Maresca, Yunker, Jones, & Bryant, 2004b). En Firmicutes y Tenericutes, se ha reportado que existe una amplia relación de las proteínas de la BC con aspectos relacionados a la patogenicidad (*Staphylococcus aureus Mu50, Mycoplasma pneumoniae)* (Liu et al., 2005) y con la resistencia a ambientes extremos (*Oceanobacillus iheyensis)* (Liang et al., 2006; Maresca et al., 2007) y

a la formación de esporas (*Bacillus indicus, B. firmus, B. altitudinis, B. Safensis*) (Takami, Takaki, & Uchiyama, 2002). En el caso de *Staphylococcus aureus*, se ha reportado que la presencia del operón CrtOPQMN, permite la formación del factor de virulencia estafiloxantina (Khaneja et al., 2010a). Se ha observado que al bloquear la carotenogénesis en *S. aureus*, se incrementa la sensibilidad a agentes oxidantes y se reduce su supervivencia en la sangre (Liu et al., 2005; Pelz et al., 2005). Para algunas de las especies de *Mycoplasma*, como *M. laidlawii*, la biosíntesis de carotenoides parece estar relacionada con el crecimiento, el aprovechamiento de la coenzima (CoA) y posible protección de la radiación solar (Razin & Rottem, 1967; Rottem et al., 1968). En el caso de la bacteria de vida libre *Bacillus cereus*, se ha descrito que algunas de las especies capaces de producir esporas, también pueden generar carotenoides (*ej.* Glicosil-apolicopeno, glicosil-4'metil-apolicopenoato), lo cual les provee de una pigmentación amarillo-anaranjada (Perez-Fons et al., 2011). Se ha comprobado que las especies de *Bacillus*, que contienen carotenoides antioxidantes como la astaxantina, son capaces de evitar que las especies reactivas generadas por la radiación UV-A, dañen sus membranas, aunque, no siempre las protege de las radiaciones UV-B o UV-C, o de sustancias como el peróxido de hidrógeno (H_2O_2) (Khaneja et al., 2010b). Para el caso del phylum Chlamydia (*Rhodopurellula baltica*), se sugiere que la presencia de carotenoides, le permite mantener su estilo de vida libre y protegerse de la radiación solar en ambientes marinos. En *Mycobacterium marinum*, se ha demostrado que la deficiencia en la biosíntesis de carotenoides incrementa la sensibilidad al peróxido de hidrógeno y reduce el crecimiento en presencia de macrófagos (Gao et al., 2003; Provvedi et al., 2008). Para Acidobacterium, se sugiere que los cetocarotenoides permiten la fotoprotección (Garcia-Costas, Graham, & Bryant, 2007).

Asimismo, se ha reportado que en las Epsilonproteobacterias (*Helicobacter pylori*), la presencia de carotenoides en su hospedero le puede ser perjudicial para su infección, es decir, la presencia de

licopeno en el intestino de ratas puede hacer que las células se recuperen de la infección por *H. pylori* (Jang, Lim, Morio, & Kim, 2012). En cuanto a la clase de Alphaproteobacteria (Rhodopseudomonas) la presencia de proteínas de la BC se relacionó con actividad de fotosíntesis (Scolnik et al., 1980). En Betaproteobacterias (*Chromobacterium violaceum*), se ha descrito que la producción del carotenoide violacelina inhibe el crecimiento del parásito *Plasmodium* sp. de humanos y ratones (Pelz et al., 2005). En el caso de las Gammaproteobacterias, la presencia de proteínas de la BC podría estar relacionada con la capacidad ser patógenos (*Photorhabdus luminescens*). En este caso, *P. luminescens*, al ser introducido por un nematodo al cuerpo de un insecto, podría requerir carotenoides para protegerse de las especies reactivas de oxígeno, que se localizan dentro del insecto (Duchaud et al., 2003; Forst, Dowds, Boemare, & Stackebrandt, 1997). Otro ejemplo de funciones de carotenoides en Gammaproteobacterias, es la bacteria *Xanthomonas axonopodis*, en la que se ha reportado que la producción de pigmentos carotenoides (xantomonadinas) tienen relevancia durante su patogénesis para protegerse de las defensas de la planta infectada (Poplawsky *et al.* 2000).

Conclusiones

El presente trabajo demuestra que los genes que codifican proteínas de la biosíntesis de carotenoides tienen una amplia y heterogénea distribución en bacterias. Lo anterior significa que la distribución no ha sido conservada a lo largo de la evolución en todos los phyla de bacterias. Además, muestra que los perfiles de proteínas generados por HMM, permiten determinar la presencia y la abundancia de la biosíntesis de carotenoides en proteomas de bacterias. Se infirió la capacidad de sintetizar carotenoides en la mayoría (133/149) de las especies de bacterias analizadas. De igual manera, se observó que las proteínas que participan en las reacciones iniciales de la biosíntesis de carotenoides fueron las más abundantes (ej CrtE, CrtB2). Al analizar la proporción del proteoma dedicada a producir proteínas de la biosíntesis de carotenoides, se encontró que es mínima (respecto al total de proteínas en el proteoma). En cuanto a la distribución, se observó que las proteínas se encuentran con mayor abundancia en los phyla con representantes capaces de realizar fotosíntesis (*ej. Cianobacteria, Chorobi*), seguido de especies con representantes infecciosos o simbiontes (*ej. Staphylococcus aureus, Mycobacterium paratuberculosis*) y por último en especies capaces de soportar condiciones extremas (ej *Bacillus*). Finalmente, la conservación heterogénea de la biosíntesis de carotenoides en bacterias puede ser debida a presiones de selección que se relacionan con los hábitos de vida (*ie.* fotosintético, patógeno, vida libre).

Perspectivas

Se espera que el presente trabajo sea una referencia para hacer análisis más amplios sobre la distribución de las proteínas de la BC en bacterias. Se considera que el estudio actual podría ser comparado con otro en que se usen secuencias semilla procedentes de bases de datos como Pfam. En este sentido, los perfiles HMM generados podrían ser mejorados al aumentar la cantidad de secuencias requeridas para entrenar dichos perfiles. Los genomas a analizar podrían ser el resto de genomas de bacterias que han sido secuenciados, incluyendo a representantes de algas, plantas y algunos hongos. Además, se podrían hacer comparaciones estructurales de los dominios conservados en los perfiles HMM y determinar los niveles de variación de las secuencias en otros organismos. De igual manera, para futuros análisis, se contempla analizar la relación entre hábitos (*ej*. patogenicidad y fotosíntesis) y la producción de carotenoides. Se contempla que podrían realizarse análisis de sintenia en un mismo género de bacteria, para analizar la conservación de la distribución de los genes de la BC, a lo largo de la evolución.

Bibliografía

A.Shehab, S., Keshk, A., & Mahgoub, H. (2012). Fast Dynamic Algorithm for Sequence Alignment Based On Bioinformatics. *International Journal of Computer Applications*, *37*(7), 54–61. http://doi.org/10.5120/4624-6636

Adler, M., Anjum, M., Berg, O. G., Andersson, D. I., & Sandegren, L. (2014). High fitness costs and instability of gene duplications reduce rates of evolution of new genes by duplication-divergence mechanisms. *Molecular Biology and Evolution*, *31*(6), 1526–35. http://doi.org/10.1093/molbev/msu111

Alberts, B., & Bray, D. (2006). *Bioquìmica* (2a ed.). Buenos Aires: Ed. Médica Panamericana,. Retrieved from http://books.google.com.mx/books?id=qrrYZJhrRm4C&printsec=frontcover&hl=es&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

Alföldi, J., & Lindblad-Toh, K. (2013). Comparative genomics as a tool to understand evolution and disease. *Cold Spring Harbor Laboratory Press*, *23*, 1063–1068. http://doi.org/10.1101/gr.157503.113.Freely

Altschul, S. F. (1998). Fundamentals of database searching. In S. Brenner & F. Lewitter (Eds.), *Trends Guide to Bioinformatics* (pp. 7–9). Cambridge,UK: Elsevier Science.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–10. http://doi.org/10.1006/jmbi.1990.9999

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST : a new generation of protein database search programs, *25*(17), 3389–3402.

Anfinsen, C. B. (1973). Principles that Govern the Folding of Protein Chains. *Science*, *181*(4096), 223–230.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., … Yeh, L.-S. L. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research*, *32*(Database issue), D115–D119. http://doi.org/10.1093/nar/gkh131

Armstrong, G. A. (1997). Genetics of eubacterial carotenoid biosynthesis: a colorful tale. *Annual Review of Microbiology*, *51*, 629–59. http://doi.org/10.1146/annurev.micro.51.1.629

Armstrong, G. A., Alberti, M., Leach, F., & Hearst, J. E. (1989). Nucleotide sequence, organization, and nature of the protein products of the carotenoid biosynthesis gene cluster of Rhodobacter capsulatus. *Mol Gen Genet*, *216*, 254–268.

Arrach, N., Schmidhauser, T. J., & Avalos, J. (2002). Mutants of the carotene cyclase domain of al-2 from Neurospora crassa. *Molecular Genetics and Genomics : MGG*, *266*(6), 914–21. http://doi.org/10.1007/s00438-001-0626-5

Assis, R., & Bachtrog, D. (2013). Neofunctionalization of young duplicate genes in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(43), 17409–14. http://doi.org/10.1073/pnas.1313759110

Attwood, T. K., & Beck, M. E. (1994). PRINTS--a protein motif fingerprint database. *Protein Engineering*, *7*(7), 841–848.

Badenhop, F., Steiger, S., Sandmann, M., & Sandmann, G. (2003). Expression and biochemical characterization of the 1-HO-carotenoid methylase CrtF from Rhodobacter capsulatus. *FEMS Microbiology Letters*, *222*(2), 237–242. http://doi.org/10.1016/S0378-1097(03)00302-1

Baldi, P., Chauvint, Y., Hunkapiller, T., & McClure, M. A. (1994). Hidden Markov models of biological primary sequence information. *Proc. Natl. Acad. Sci. USA*, *91*(February), 1059–1063.

Barredo, L. (2012). *Microbial Carotenoids from Bacteria and Microalgae*. (J.-L. Barredo, Ed.) (Vol. 892). Totowa, NJ: Springer Science+Busines Media B.V. http://doi.org/10.1007/978-1-61779-879-5

Baum, L. E., & Eagon, J. a. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, *73*(3), 360–364. http://doi.org/10.1090/S0002-9904-1967-11751-8

Bautista, J. a, Rappaport, F., Guergova-Kuras, M., Cohen, R. O., Golbeck, J. H., Wang, J. Y., … Diner, B. a. (2005). Biochemical and biophysical characterization of photosystem I from phytoene desaturase and zeta-carotene desaturase deletion mutants of Synechocystis Sp. PCC 6803: evidence for PsaA- and PsaB-side electron transport in cyanobacteria. *The Journal of Biological Chemistry*, *280*(20), 20030–41. http://doi.org/10.1074/jbc.M500809200

Blumenthal, T. (2004). Operons in eukaryotes. *Briefings in Functional Genomics & Proteomics*, *3*(3), 199–211. http://doi.org/10.1093/bfgp/3.3.199

Bork, P., & Koonin, E. V. (1996). Protein sequence motifs. *Current Opinion in Structural Biology*, *6*(3), 366–376.

Brenner, S. E. (2001). A tour of structural genomics. *Nature Reviews. Genetics*, *2*(10), 801–809. http://doi.org/10.1038/35093574

Britton, G., Armitt, G. M., Lau, S. Y. M., Patel, A. K., & Shone, C. C. (1982). *Carotenoid Chemistry and Biochemistry. Carotenoid Chemistry and Biochemistry*. Elsevier. http://doi.org/10.1016/B978-0-08-026224-6.50020-6

Brown, T. A. (2002). *Genomes* (2nd ed.). Manchester, UK: Wiley-Liss.

Bykova, N. a, Favorov, A. V, & Mironov, A. a. (2013). Hidden Markov models for evolution and comparative genomics analysis. *PloS One*, *8*(6), e65012. http://doi.org/10.1371/journal.pone.0065012

Choi, S.-K., Matsuda, S., Hoshino, T., Peng, X., & Misawa, N. (2006). Characterization of bacterial beta-carotene 3,3'-hydroxylases, CrtZ, and P450 in astaxanthin biosynthetic pathway and adonirubin production by gene combination in Escherichia coli. *Applied Microbiology and Biotechnology*, *72*(6), 1238–1246. http://doi.org/10.1007/s00253-006-0426-2

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B., & Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science (New York, N.Y.)*, *311*(5765), 1283–7. http://doi.org/10.1126/science.1123061

Coffin, J. M., Hughes, S. H., & Varmus, H. E. (1997). Retroviruses -- NCBI Bookshelf. *Retroviruses*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press. http://doi.org/ISBN-10: 0-87969-571-4; Bookshelf ID: NBK19376

Cogdell, R. J., Howard, T. D., Bittl, R., Schlodder, E., Geisenheimer, I., & Lubitz, W. (2000). How carotenoids protect bacterial photosynthesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *355*(1402), 1345–9. http://doi.org/10.1098/rstb.2000.0696

Crick, F. H. C. (1956). On protein synthesis. *Symposia of the Society for Experimental Biology*, *XII*, 139–163. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/13580867

Cunningham, F. X., & Gantt, E. (1998). Genes and Enzymes of Carotenoid Biosynthesis in Plants. *Annual Review of Plant Physiology and Plant Molecular Biology*, *49*, 557–583. http://doi.org/10.1146/annurev.arplant.49.1.557

Cunningham, F. X., Pogson, B. J., Sun, Z., McDonald, K. A., DellaPenna, D., & Gantt, E. (1996). Functional analysis of the beta and epsilon lycopene cyclase enzymes of arabidopsis reveals a mechanism for control of cyclic carotenoid formation. *The Plant Cell*, *8*(September), 1613–1626.

Cunningham, F. X., Sun, Z., Chamovitz, D., Hirschberg, J., & Gantt, E. (1994). Molecular Structure and Enzymatic Function of Lycopene Cyclase from the Cyanobacterium. *The Plant Cell*, *6*(August), 1107–1121.

D'Haene, S. E., Crouch, L. I., Jones, M. R., & Frese, R. N. (2014). Organization in photosynthetic membranes of purple bacteria in vivo: The role of carotenoids. *Biochimica et Biophysica Acta - Bioenergetics*, *1837*(10), 1665–1673. http://doi.org/10.1016/j.bbabio.2014.07.003

Da Silva-Mendes, A. F., Fontes-Soares, V. L., & Cardoso-Costa, M. G. (2015). Carotenoid Biosynthesis Genomics. In C. Chen (Ed.), *Pigments in Fruit and Vegetables* (pp. 9–19). New York, NY:

Springer Science+Busines Media B.V. http://doi.org/10.1007/978-1-4939-2356-4

Dai, J., & Cheng, J. (2008). HMMEditor: a visual editing tool for profile hidden Markov model. *BMC Genomics*, *9 (Suppl 1*(S8), 1–7. http://doi.org/10.1186/1471-2164-9-S1-S8

Das, A., Yoon, S. H., Lee, S. H., Kim, J. Y., Oh, D. K., & Kim, S. W. (2007). An update on microbial carotenoid production: Application of recent metabolic engineering tools. *Applied Microbiology and Biotechnology*, *77*(3), 505–512. http://doi.org/10.1007/s00253-007-1206-3

Daubin, V., Gouy, M., & Perrière, G. (2002a). A Phylogenomic Approach to Bacterial Phylogeny : Evidence of a Core of Genes Sharing a Common History. *Genome Research*, *12*, 1080–1090. http://doi.org/10.1101/gr.187002.base

Daubin, V., Gouy, M., & Perrière, G. (2002b). A phylogenomic approach to bacterial phylogeny: Evidence of a core of genes sharing a common history. *Genome Research*, *12*(7), 1080–1090. http://doi.org/10.1101/gr.187002

Dayhoff, M., & Schwartz, R. (1978). A Model of Evolutionary Change in Proteins. *In Atlas of Protein Sequence and Structure*, *5*, 345–352. http://doi.org/10.1.1.145.4315

Devlin, T. M. (2004). *Bioquímica: libro de texto con aplicaciones clínicas* (4th ed.). España: Reverté. Retrieved from https://books.google.com.mx/books?id=p3DCb9lTLx8C

Dieser, M., Greenwood, M., & Foreman, C. M. (2010). Carotenoid Pigmentation in Antarctic Heterotrophic Bacteria as a Strategy to Withstand Environmental Stresses. *Arctic, Antarctic, and Alpine Research*, *42*(4), 396–405. http://doi.org/10.1657/1938-4246-42.4.396

Dorcas, J., & Orengo, F. (2013). *Fundamentos de biología molecular*. Barcelona a: UOC (Universitat Oberta de Catalunya).

Duchaud, E., Rusniok, C., Frangeul, L., Buchrieser, C., Givaudan, A., Taourit, S., … Kunst, F. (2003). The genome sequence of the entomopathogenic bacterium Photorhabdus luminescens. *Nature Biotechnology*, *21*(11), 1307–13. http://doi.org/10.1038/nbt886

Eddy, S. R. (1998). Profile hidden Markov models. *Bionformatics Review*, *14*(9), 755–763.

Eddy, S. R. (2008). A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Computational Biology*, *4*(5), e1000069. http://doi.org/10.1371/journal.pcbi.1000069

Eddy, S. R. (2010). HMMER User ' s Guide. United States of America: Howard Hughes Medical Institute. Retrieved from ftp://selab.janelia.org/pub/software/hmmer/CURRENT/Userguide.pdf

Edwards, D. J., & Holt, K. E. (2013). Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microbial Informatics and Experimentation*, *3*(1), 2. http://doi.org/10.1186/2042-5783-3-2

Eisen, J. A., & Fraser, C. M. (2003). Phylogenomics: Intersection of Evolution and Genomics. *Science*, *300*, 1706–1708.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., … Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Research*, *42*(Database issue), D222–30. http://doi.org/10.1093/nar/gkt1223

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, *39*(Web Server issue), W29–37. http://doi.org/10.1093/nar/gkr367

Forst, S., Dowds, B., Boemare, N., & Stackebrandt, E. (1997). Xenorhabdus and Photorhabdus spp.: Bugs That Kill Bugs. *Annu. Rev. Microbiol*, *51*, 47–72.

Fraser, C. M., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R., Lathigra, R., … Venter, J. C. (1997). Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi. *Nature*, *390*(6660), 580–586. http://doi.org/10.1038/37551

Frenz, C. M. (2008). Introduction to Searching with Regular Expressions. In *Proceedings of the 2008 Trenton Computer Festival*. Retrieved from http://arxiv.org/abs/0810.1732v1\npapers2://publication/uuid/E0CCE658-EB35-40B6-BE94-6D03F98D41E5

Frigaard, N.-U., Maresca, J. a, Yunker, C. E., Jones, A. D., & Bryant, D. A. (2004a). Genetic manipulation of carotenoid biosynthesis in the green sulfur bacterium Chlorobium tepidum. *Journal of Bacteriology*, *186*(16), 5210–20. http://doi.org/10.1128/JB.186.16.5210-5220.2004

Frigaard, N.-U., Maresca, J. a, Yunker, C. E., Jones, A. D., & Bryant, D. A. (2004b). Genetic manipulation of carotenoid biosynthesis in the green sulfur bacterium Chlorobium tepidum. *Journal of Bacteriology*, *186*(16), 5210–20. http://doi.org/10.1128/JB.186.16.5210-5220.2004

Furubayashi, M., Li, L., Katabami, A., Saito, K., & Umeno, D. (2014). Construction of carotenoid biosynthetic pathways using squalene synthase. *FEBS Letters*, *588*(3), 436–42. http://doi.org/10.1016/j.febslet.2013.12.003

Gallagher, C. E., Matthews, P. D., Li, F., & Wurtzel, E. T. (2004). Gene duplication in the carotenoid biosynthetic pathway preceded evolution of the grasses. *Plant Physiology*, *135*(3), 1776–1783. http://doi.org/10.1104/pp.104.039818

Gao, L. Y., Groger, R., Cox, J. S., Beverley, S. M., Lawson, E. H., & Brown, E. J. (2003). Transposon Mutagenesis of Mycobacterium marinum Identifies a Locus Linking Pigmentation and Intracellular Survival. *Infection and Immunity*, *71*(2), 922–929. http://doi.org/10.1128/IAI.71.2.922-929.2003

Garcia-Costas, A. M., Graham, J. E., & Bryant, D. A. (2007). Ketocarotenoids in chlorosomes of the acidobacterium Candidatus Chloracidobacterium thermophilum. In J. Allen, E. Gantt, J. H. Golbeck, & B. Osmond (Eds.), *Proceedings of the XIVth International Congress on Photosynthesis, Glasgow, Scotland, July 22-27, 2007* (pp. 1–7). Netherlands: Springer Netherlands.

Gaudriault, S., Duchaud, E., Lanois, A., Canoy, A. S., Bourot, S., DeRose, R., … Givaudan, A. (2006). Whole-genome comparison between Photorhabdus strains to identify genomic regions involved in the specificity of nematode interaction. *Journal of Bacteriology*, *188*(2), 809–814. http://doi.org/10.1128/JB.188.2.809-814.2006

Gerjets, T., Steiger, S., & Sandmann, G. (2009). Catalytic properties of the expressed acyclic carotenoid 2-ketolases from Rhodobacter capsulatus and Rubrivivax gelatinosus. *Biochimica et Biophysica Acta*, *1791*(2), 125–31. http://doi.org/10.1016/j.bbalip.2008.12.006

Gómez-García, M. D. R., & Ochoa-Alejo, N. (2013). Biochemistry and molecular Biology of carotenoid biosynthesis in chili peppers (Capsicum spp.). *International Journal of Molecular Sciences*, *14*(9), 19025–19053. http://doi.org/10.3390/ijms140919025

Goodwin, T. W. (1972). Carotenoids in fungi and non-photosynthetic bacteria. *Progress in Industrial Microbiology*, *11*, 29–88.

Goodwin, T. W. (1980). Fungi. In T. W. Goodwin (Ed.), *The Biochemistry of the Carotenoids* (pp. 257–290). Netherlands: Springer Netherlands.

Gottlieb, K., Albermann, C., & Sprenger, G. A. (2014). Improvement of L-phenylalanine production from glycerol by recombinant Escherichia coli strains: the role of extra copies of glpK, glpX, and tktA genes. *Microbial Cell Factories*, *13*(1), 96. http://doi.org/10.1186/s12934-014-0096-1

Griswold, A. (2008). Genome Packaging in Prokaryotes : the Circular Chromosome of E coli. *Nature Education*, *1*(1), 1–6. Retrieved from http://www.nature.com/scitable/topicpage/genome-packaging-in-prokaryotes-the-circular-chromosome-9113

Gupta, R. S., & Lorenzini, E. (2007). Phylogeny and molecular signatures (conserved proteins and indels) that are specific for the Bacteroidetes and Chlorobi species. *BMC Evolutionary Biology*, *7*(71), 1–18. http://doi.org/10.1186/1471-2148-7-71

Han, K., Li, Z., Peng, R., Zhu, L., Zhou, T., Wang, L., … Li, Y. (2013). Extraordinary expansion of a Sorangium cellulosum genome from an alkaline milieu. *Scientific Reports*, *3*, 2101. http://doi.org/10.1038/srep02101

Hardison, R. C. (2003). Comparative genomics. *PLoS Biology*, *1*(2), 156–160. http://doi.org/10.1371/journal.pbio.0000058

Hashtroudi, M. S., Shariatmadari, Z., Riahi, H., & Ghassempour, A. (2013). Analysis of Anabaena vaginicola and Nostoc calcicola from Northern Iran, as rich sources of major carotenoids. *Food Chemistry*, *136*(3-4), 1148–53. http://doi.org/10.1016/j.foodchem.2012.09.055

Heider, S. A. E., Peters-Wendisch, P., & Wendisch, V. F. (2012). Carotenoid biosynthesis and overproduction in Corynebacterium glutamicum. *BMC Microbiology*, *12*(1), 198. http://doi.org/10.1186/1471-2180-12-198

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(22), 10915–10919. http://doi.org/10.1073/pnas.89.22.10915

Herbert, K. G., Spirollari, J., Wang, J. T. L., Piel, W. H., Westbrook, J., Barker, W. C., … Wu, C. H. (2008). Bioinformatic Databases. In B. Wah (Ed.), *Wiley Encyclopedia of Computer Science and Engineering* (pp. 1–10). United States of America: John Wiley & Sons.

Hirschberg, J., Cohen, M., Harker, M., Lotan, T., Mann, V., & Pecker, I. (1997). Molecular genetics of the carotenoid biosynthesis pathway in plants and algae. *Pure & Appl. Chem*, *69*(10), 2151–2158.

Hiseni, A., Arends, I. W. C. E., & Otten, L. G. (2011). Biochemical characterization of the carotenoid 1,2-hydratases (CrtC) from Rubrivivax gelatinosus and Thiocapsa roseopersicina. *Applied Microbiology and Biotechnology*, *91*(4), 1029–36. http://doi.org/10.1007/s00253-011-3324-1

Hoberman, R., & Durand, D. (2011). HMM Lecture Notes HMM topology and Profile HMMs. *Computational Genomics and Molecular Biology,* 1–9.

Hogeweg, P. (2011). The Roots of Bioinformatics in Theoretical Biology. *PLoS Computational Biology*, *7*(3), 1–5. http://doi.org/10.1371/journal.pcbi.1002021

Hoshino, T., Fujii, R., & Nakahara, T. (1993). Molecular Cloning and Sequence Analysis of the crtB Gene of Thermus thermophilus HB27 , an Extreme Thermophile Producing Carotenoid Pigments. *Appl. Environ. Microbiol.*, *59*(9), 3150–3153.

Huang, J. Y., & Brutlag, D. L. (2001). The EMOTIF database. *Nucleic Acids Research*, *29*(1), 202–204. http://doi.org/10.1093/nar/29.1.202

Iniesta, A. A., Cervantes, M., & Murillo, F. J. (2007). Cooperation of two carotene desaturases in the production of lycopene in Myxococcus xanthus. *The FEBS Journal*, *274*(16), 4306–14. http://doi.org/10.1111/j.1742-4658.2007.05960.x

IUPAC-IUB Comission on Biochemical Nomenclature. (1974). Nomenclature of multiple forms of enzymes. *Pure & Appl. Chem*, *40*(309), 310–314.

Jacob, F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, *3*, 318–356.

Jang, S. H., Lim, J. W., Morio, T., & Kim, H. (2012). Lycopene inhibits Helicobacter pylori-induced ATM/ATR-dependent DNA damage response in gastric epithelial AGS cells. *Free Radical Biology & Medicine*, *52*(3), 607–15. http://doi.org/10.1016/j.freeradbiomed.2011.11.010

Janin, J., & Chothia, C. (1985). Domains in proteins: Definitions, location, and structural principles. In B. T.-M. in Enzymology (Ed.), *Diffraction Methods for Biological Macromolecules Part B* (Vol. Volume 115, pp. 420–430). Academic Press. http://doi.org/http://dx.doi.org/10.1016/0076-6879(85)15030-5

Jáuregui-Carranco, E. M., Calvo-Carrillo, M. de la C., & Pérez-Gil-Romo, F. (2011). Carotenoides y su función antioxidante : Revisión. *Archivos Latinoamericanos de Nutrición*, *61*(3), 233–241.

Jeon, B. Y., Kim, B. Y., Jung, I. L., & Park, D. H. (2012). Metabolic roles of carotenoid produced by non-photosynthetic bacterium Gordonia alkanivorans SKF120101. *Journal of Microbiology and Biotechnology*, *22*(11), 1471–1477. http://doi.org/10.4014/jmb.1207.07038

Jones, D. T. (1999). Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. *J. Mol. Biol.*, 195–202. http://doi.org/10.1111/j.1464-410X.2007.07404.x

Karplus, K., Barrett, C., & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, *14*(10), 846–856. http://doi.org/10.1093/bioinformatics/14.10.846

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. http://doi.org/10.1093/nar/gkf436

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–80. http://doi.org/10.1093/molbev/mst010

Katoh, K., & Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*, *9*(4), 286–98. http://doi.org/10.1093/bib/bbn013

Kaur, H., Singh, A., & Singh, P. (2008). Comparison of Variants of BLAST. *Proceedings of the International MultiConference of Engineers and Computer Scientists*, *1*, 19–21. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.148.5891

Khaneja, R., Perez-Fons, L., Fakhry, S., Baccigalupi, L., Steiger, S., To, E., … Cutting, S. M. (2010a). Carotenoids found in Bacillus. *Journal of Applied Microbiology*, *108*(6), 1889–902. http://doi.org/10.1111/j.1365-2672.2009.04590.x

Khaneja, R., Perez-Fons, L., Fakhry, S., Baccigalupi, L., Steiger, S., To, E., … Cutting, S. M. (2010b). Carotenoids found in Bacillus. *Journal of Applied Microbiology*, *108*(6), 1889–1902. http://doi.org/10.1111/j.1365-2672.2009.04590.x

Klassen, J. L. (2010). Phylogenetic and evolutionary patterns in microbial carotenoid biosynthesis are revealed by comparative genomics. *PloS One*, *5*(6), e11257. http://doi.org/10.1371/journal.pone.0011257

Kleppe, K., Steinar, Ö., & Lossius, I. (1979). The bacterial nucleoid revisited. *Journal of General Microbiology*, *112*, 1–13. http://doi.org/10.1099/00221287-112-1-1

Koonin, E. V., & Wolf, Y. I. (2008). Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world. *Nucleic Acids Research*, *36*(21), 6688–6719. http://doi.org/10.1093/nar/gkn668

Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M. R., Ahn, T.-H., … Ussery, D. W. (2015). Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics*, *15*, 141–161. http://doi.org/10.1007/s10142-015-0433-4

Lang, H. P., Cogdell, R. J., Gardiner, A. T., & Hunter, N. (1994). Early Steps in Carotenoid Biosynthesis: Sequences and Transcriptional Analysis of the crtI and crtB Genes of Rhodobacter sphaeroides and Overexpression and Reactivation of crtI in Escherichia coli and R. sphaeroides. *Journal of Bacteriology*, *176*(13), 3859–3869.

Lange, B. M., Rujan, T., Martin, W., & Croteau, R. (2000). Isoprenoid biosynthesis: the evolution of two ancient and distinct pathways across genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(24), 13172–7. http://doi.org/10.1073/pnas.240454797

Letunic, I., & Bork, P. (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Research*, *39*(Web Server issue), W475–8. http://doi.org/10.1093/nar/gkr201

Liang, C., Zhao, F., Wei, W., Wen, Z., & Qin, S. (2006). Carotenoid Biosynthesis in Cyanobacteria : Structural and Evolutionary Scenarios Based on Comparative Genomics. *Int. J. Biol. Sci.*, *2*(4), 197–207.

Liu, G. Y., Essex, A., Buchanan, J. T., Datta, V., Hoffman, H. M., Bastian, J. F., … Nizet, V. (2005). Staphylococcus aureus golden pigment impairs neutrophil killing and promotes virulence through its antioxidant activity. *The Journal of Experimental Medicine*, *202*(2), 209–15. http://doi.org/10.1084/jem.20050846

Lodish, H., Berk, A., Zipursky, L., Matsudaira, P., Baltimore, D., & Darnell, J. E. (2000). *Molecular Cell Biology* (4th ed.). New York, NY: W. H. Freeman and Company. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK21475/?term=genome

Logan, B., Moreno, P., Suzek, B., Weng, Z., & Kasif, S. (2001). *A Study of remote homology detection*. *Technical Report Series* (Vol. CRL 2001/0). Cambridge,UK. http://doi.org/10.1017/CBO9781107415324.004

Lohr, M., Im, C.-S., & Grossman, A. (2005). Genome-Based Examination of Chlorophyll and Carotenoid Biosynthesis in Chlamydomonas reinhardtii 1 [w]. *Plant Physiology*, *138*(May), 490–515. http://doi.org/10.1104/pp.104.056069.490

Lopes, S. C. P., Blanco, Y. C., Justo, G. Z., Nogueira, P. a, Rodrigues, F. L. S., Goelnitz, U., … Costa, F. T. M. (2009). Violacein extracted from Chromobacterium violaceum inhibits Plasmodium growth in vitro and in vivo. *Antimicrobial Agents and Chemotherapy*, *53*(5), 2149–52. http://doi.org/10.1128/AAC.00693-08

López-López, M., López-Gutiérrez, A. U., Sainz-Espuñes, T. del R., & Rosales-Torres, A. M. (2005). ¿Qué sabe usted acerca de...Genómica? *Revista Mexicana de Ciencias Farmacéuticas*, *36(1)*, 42–44.

Luscombe, N. M., Greenbaum, D., & Gerstein, M. (2001). What is bioinformatics ? An introduction and overview. *Yearbook of Medical Informatics*, 83–100.

Malpica, M., Fraile, A., Moreno, I., Obies, C. I., Drake, J. W., & Garcı, F. (2002). The Rate and Character of Spontaneous Mutation in an RNA Virus. *Genetics*, *162*(December), 1505–1511.

Maresca, J. A., & Bryant, D. A. (2006). Two Genes Encoding New Carotenoid-Modifying Enzymes in the Green Sulfur Bacterium Chlorobium tepidum. *Journal of Bacteriology*, *109*(17), 6217–6223. http://doi.org/10.1128/JB.00766-06

Maresca, J. A., Graham, J. E., Wu, M., Eisen, J. A., & Bryant, D. A. (2007). Identification of a fourth family of lycopene cyclases in photosynthetic bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(28), 11784–11789. http://doi.org/10.1073/pnas.0702984104

Martin, D., Procter, J., Waterhouse, A., Shehata, S., & Barton, G. (2013). Jalview 2.8, (January).

Masamoto, K., Misawa, N., Kaneko, T., Kikuno, R., & Toh, H. (1998). Beta -Carotene Hydroxylase Gene from the Cyanobacterium Synechocystis sp . PCC6803. *Plant Cell Physiol*, *39*(5), 560–564.

Mathews, M. M., & Krinsky, N. I. (1965). THE RELATIONSHIP BETWEEN CAROTENOID PIGMENTS AND RESISTANCE TO RADIATION IN NON-PHOTOSYNTHETIC BACTERIA. *Photochemistry and Photobiology*, *4*(4), 813–817. http://doi.org/10.1111/j.1751-1097.1965.tb07923.x

Mathews, M. M., & Sistrom, W. R. (1959). Function of carotenoid pigments in non-photosynthetic bacteria. *Nature*, *184(Suppl* , 1892–1893. http://doi.org/10.1038/1841892a0

Meléndez-Martínez, A. J., Vicario, I. M., & Heredia, F. J. (2007). Pigmentos carotenoides : consideraciones estructurales y fisicoquímicas. *Archivos Latinoamericanos de Nutrición*, *57*, 109–117.

Misawa, N., Nakagawa, M., Kobayashi, K., Yamano, S., Izawa, Y., Nakamura, K., & Harashima, K. (1990). Elucidation of the Erwinia uredovora Carotenoid Biosynthetic Pathway by Functional Analysis of Gene Products Expressed in Escherichia coli. *Journal of Bacteriology*, *172*(12), 6704–6712.

Moeller, R., Horneck, G., Facius, R., & Stackebrandt, E. (2005). Role of pigmentation in protecting Bacillus sp. endospores against environmental UV radiation. *FEMS Microbiology Ecology*, *51*(2), 231–6. http://doi.org/10.1016/j.femsec.2004.08.008

Mona, S., & Parker, J. (1999). Introduction to Hidden Markov Models. Retrieved from https://www.cs.princeton.edu/~mona/Lecture/HMM1.pdf

Mulkidjanian, A. Y., & Galperin, M. Y. (2013). Chapter One – A Time to Scatter Genes and a Time to Gather Them: Evolution of Photosynthesis Genes in Bacteria. In *BS:ABR* (Vol. 66, pp. 1–35). Elsevier. http://doi.org/10.1016/B978-0-12-397923-0.00001-1

Mulkidjanian, A. Y., Koonin, E. V, Makarova, K. S., Mekhedov, S. L., Sorokin, A., Wolf, Y. I., … Galperin, M. Y. (2006). The cyanobacterial genome core and the origin of photosynthesis. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(35), 13126–31. http://doi.org/10.1073/pnas.0605709103

Müller-Esterl, W. (2009). *Bioquímica. Fundamentos para Medicina y Ciencias de la Vida*. Reverte. Retrieved from http://books.google.com/books?id=X2YVG6Fzp1UC&pgis=1

Nahas, M. El, Kassim, S., & Shikoun, N. (2012). Profile Hidden Markov Model for Detection and Prediction of Hepatitis C Virus Mutation. *International Journal of Computer Science Issues*, *9*(5), 251–256.

National Center for Biotechnology Information. (2014). NCBI Prokaryotic Genome Annotation Pipeline. Retrieved from http://www.ncbi.nlm.nih.gov/genome/annotation_prok/

Neddleman, S. B., & Wunsch, C. D. (1970). A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.*, *48*, 443–453.

Nelson, D. L., & Cox, M. M. (2008a). *Lehninger Principles of biochemistry* (Fifth edit). New York, NY: W. H. Freeman and Company.

Nelson, D. L., & Cox, M. M. (2008b). *Lehninger Principles of Biochemistry* (5th ed.). New York, NY: W. H. Freeman and Company.

Owens, M. (2015). Inkscape. Retrieved July 23, 2015, from https://inkscape.org/en/

Paiva, S. A. R., & Russell, R. M. (1999). β-Carotene and Other Carotenoids as Antioxidants. *Journal of the American College of Nutrition*, *18*(5), 426–433. http://doi.org/10.1080/07315724.1999.10718880

Pearson, W. R. (2013). An Introduction to Sequence Similarity ("Homology") Searching. *Current Protocols in Bioinformatics*, 1–8. http://doi.org/10.1002/0471250953.bi0301s42

Pellicer, J., Fay, M. F., & Leitch, I. J. (2010). The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society*, *164*(1), 10–15. http://doi.org/10.1111/j.1095-8339.2010.01072.x

Pelz, A., Wieland, K.-P., Putzbach, K., Hentschel, P., Albert, K., & Götz, F. (2005). Structure and biosynthesis of staphyloxanthin from Staphylococcus aureus. *The Journal of Biological Chemistry*, *280*(37), 32493–8. http://doi.org/10.1074/jbc.M505070200

Perez-Fons, L., Steiger, S., Khaneja, R., Bramley, P. M., Cutting, S. M., Sandmann, G., & Fraser, P. D. (2011). Identification and the developmental formation of carotenoid pigments in the yellow/orange Bacillus spore-formers. *Biochimica et Biophysica Acta*, *1811*(3), 177–85. http://doi.org/10.1016/j.bbalip.2010.12.009

Phadwal, K. (2005). Carotenoid biosynthetic pathway: molecular phylogenies and evolutionary behavior of crt genes in eubacteria. *Gene*, *345*(1), 35–43. http://doi.org/10.1016/j.gene.2004.11.038

Pietrokovski, S., Henikoff, J. G., & Henikoff, S. (1996). The Blocks database--a system for protein classification. *Nucleic Acids Research*, *24*(1), 197–200. http://doi.org/5s1011 [pii]

Prada-Alonso, S. (2013). *Cadenas de Markov en la Investigación del Genoma*. Universidadde de Vigo, Universidade de Santigo de Compostella, Universidadde da Coruña.

Provvedi, R., Kocíncová, D., Donà, V., Euphrasie, D., Daffé, M., Etienne, G., … Reyrat, J.-M. (2008). SigF controls carotenoid pigment production and affects transformation efficiency and hydrogen peroxide sensitivity in Mycobacterium smegmatis. *Journal of Bacteriology*, *190*(23), 7859–63. http://doi.org/10.1128/JB.00714-08

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., & Lopez, R. (2005). InterProScan: protein domains identifier. *Nucleic Acids Research*, *33*(Web Server issue), W116–20. http://doi.org/10.1093/nar/gki442

R Core Team. (2013a, February). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://doi.org/10.11120/msor.2001.01010023

R Core Team. (2013b, February). R: A language and environment for statistical computing. http://doi.org/10.11120/msor.2001.01010023

Razin, S., & Rottem, S. (1967). Role of Carotenoids and Cholesterol in the Growth of Mycoplasma laidlawii. *Journal of Bacteriology*, *93*(3), 1181–1182.

Rehm, B. (2001). Bioinformatic tools for DNA/protein sequence analysis, functional assignment of

genes and protein classification. *Applied Microbiology and Biotechnology*, *57*(5-6), 579–592. http://doi.org/10.1007/s00253-001-0844-0

Richardson, J. S. (1981). The anatomy & taxonomy of protein structure. *Advances in Protein Chemistry*, *34*, 1–131. http://doi.org/10.1016/S0022-2836(77)80048-X

Rodríguez Villalón, A. (2010). *Biosíntesis de carotenoides en Escherichia coli y tejidos no fotosintéticos de Arabidopsis thaliana*. Universitat de Barcelona.

Rosa-Putra, S., Disch, A., Bravo, J.-M., & Rohmer, M. (1998). Distribution of mevalonate and glyceraldehyde 3-phosphate/pyruvate routes for isoprenoid biosynthesis in some Gram-negative bacteria and mycobacteria. *FEMS Microbiology Letters*, *164*(1), 169–175. http://doi.org/10.1111/j.1574-6968.1998.tb13082.x

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering*, *12*(2), 85–94. http://doi.org/10.1093/protein/12.2.85

Rottem, S., Gottfried, L., & Razin, S. (1968). Carotenoids as protectors against photodynamic inactivation of the adenosine trifosfatasa of Mycoplasma laidlawii membranes. *Biochem. J.1*, *109*, 707–708.

Schneider, C., Böger, P., & Sandmann, G. (1997). Phytoene Desaturase : Heterologous Expression in an Active State , Purification , and Biochemical Properties. *Protein Expression and Purification*, *10*, 175–179.

Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, *18*(20), 6097–6100. http://doi.org/10.1093/nar/18.20.6097

Schuster-böckler, B., Schultz, J., & Rahmann, S. (2004). HMM Logos for visualization of protein families. *BMC Bioinformatics*, *5*(7), 1–8.

Scolnik, P. A., Walker, M. A., & Marrs, B. L. (1980). Biosynthesis of Carotenoids Derived from Neurosporene in Rhodopseudomonas capsulata. *The Journal of Biological Chemistry*, *255*(6), 2427–2432.

Sentausa, E., & Fournier, P. E. (2013). Advantages and limitations of genomics in prokaryotic taxonomy. *Clinical Microbiology and Infection*, *19*(9), 790–795. http://doi.org/10.1111/1469-0691.12181

Shiryev, S. A., Papadopoulos, J. S., Schäffer, A. A., & Agarwala, R. (2007). Improved BLAST searches using longer words for protein seeding. *Bioinformatics (Oxford, England)*, *23*(21), 2949–2951. http://doi.org/10.1093/bioinformatics/btm479

Shivanand, M. K., Hearst, J. E., & Poulter, C. D. (1992). The crtE gene in Erwinia herbicola encodes geranylgeranyl diphosphate synthase. *Proceedings of the National Academy of Sciences of the*

United States of America, *89*, 6761–6764.

Sieiro, C., Poza, M., de Miguel, T., & Villa, T. G. (2003). Genetic basis of microbial carotenogenesis. *Int. Microbiol*, *6*(1), 11–6. http://doi.org/10.1007/s10123-003-0097-0

Sigrist, C. J. a, Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., … Bucher, P. (2002). PROSITE: a documented database using patterns and profiles as motif descriptors. *Briefings in Bioinformatics*, *3*(3), 265–274. http://doi.org/10.1093/bib/3.3.265

Sivashankari, S., & Shanmughavel, P. (2007). Comparative genomics - a perspective. *Bioinformation*, *1*(9), 376–378. http://doi.org/10.6026/97320630001376

Šlouf, V., Chábera, P., Olsen, J. D., Martin, E. C., Qian, P., Hunter, C. N., & Polívka, T. (2012). Photoprotection in a purple phototrophic bacterium mediated by oxygen-dependent alteration of carotenoid excited-state properties. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(22), 8570–5. http://doi.org/10.1073/pnas.1201413109

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Molecular Biology*, *147*(1), 195–197. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7265238

Steiger, S., Mazet, A., & Sandmann, G. (2003). Heterologous expression, purification, and enzymatic characterization of the acyclic carotenoid 1,2-hydratase from Rubrivivax gelatinosus. *Archives of Biochemistry and Biophysics*, *414*(1), 51–58. http://doi.org/10.1016/S0003-9861(03)00099-7

Sudha, M. P. (2014). Sequence Alignment in DNA Using Smith Waterman and Needleman Algorithms, *5*(4), 5957–5960.

Takaichi, S. (2009). Chapter 6 Distribution and Biosynthesis of Carotenoids. In C. N. Hunter, F. Daldal, M. C. Thurnauer, & J. T. Beatty (Eds.), *The Purple Phototrophic Bacteria* (pp. 97–117). Netherlands: Springer Netherlands.

Takaichi, S. (2011). Carotenoids in algae: distributions, biosyntheses and functions. *Marine Drugs*, *9*(6), 1101–18. http://doi.org/10.3390/md9061101

Takaichi, S., Maoka, T., & Masamoto, K. (2001). Myxoxanthophyll in Synechocystis sp . PCC 6803 is Myxol 2’-Dimethyl- Fucoside, (3R, 2'S) -Myxol 2'- (2,4-di-O-Methyl- alfa- L -Fucoside), not Rhamnoside. *Plant Cell Physiol*, *42*(7), 756–762.

Takami, H., Takaki, Y., & Uchiyama, I. (2002). Genome sequence of Oceanobacillus iheyensis isolated from the Iheya Ridge and its unexpected adaptive capabilities to extreme environments. *Nucleic Acids Research*, *30*(18), 3927–3935.

Takano, H., Obitsu, S., Beppu, T., & Ueda, K. (2005). Light-induced carotenogenesis in Streptomyces coelicolor A3(2): identification of an extracytoplasmic function sigma factor that directs photodependent transcription of the carotenoid biosynthesis gene cluster. *Journal of Bacteriology*,

187(5), 1825–32. http://doi.org/10.1128/JB.187.5.1825-1832.2005

Tao, L., Schenzle, A., Odom, J. M., & Cheng, Q. (2005). Novel Carotenoid Oxidase Involved in Biosynthesis of 4,4'-Diapolycopene Dialdehyde. *Biological and Chemical Sciences and Engineering*, *71*(6), 3294–3301. http://doi.org/10.1128/AEM.71.6.3294

Tóth, T. N., Chukhutsina, V., Domonkos, I., Knoppová, J., Komenda, J., Kis, M., … van Amerongen, H. (2015). Carotenoids are essential for the assembly of cyanobacterial photosynthetic complexes. *Biochimica et Biophysica Acta*, *1847*(10), 1153–1165. http://doi.org/10.1016/j.bbabio.2015.05.020

Tsuchiya, T., Takaichi, S., Misawa, N., Maoka, T., Miyashita, H., & Mimuro, M. (2005). The cyanobacterium Gloeobacter violaceus PCC 7421 uses bacterial-type phytoene desaturase in carotenoid biosynthesis. *FEBS Letters*, *579*(10), 2125–2129. http://doi.org/10.1016/j.febslet.2005.02.066

Umeno, D., Tobias, A. V, & Arnold, F. H. (2002). Evolution of the C 30 Carotenoid Synthase CrtM for Function in a C 40 Pathway. *Journal of Bacteriology*, *184*(23), 6690–6699. http://doi.org/10.1128/JB.184.23.6690

Umeno, D., Tobias, A. V, & Frances, H. (2005). Diversifying Carotenoid Biosynthetic Pathways by Directed Evolution Diversifying Carotenoid Biosynthetic Pathways by Directed Evolution. *Microbiol. Mol. Biol. Rev.*, *69*(1), 51–78. http://doi.org/10.1128/MMBR.69.1.51

van Iersel, M. P., Kelder, T., Pico, A. R., Hanspers, K., Coort, S., Conklin, B. R., & Evelo, C. (2008). Presenting and exploring biological pathways with PathVisio. *BMC Bioinformatics*, *9*, 399. http://doi.org/10.1186/1471-2105-9-399

Venugopal, K. R., Srinivasa, K. G., & Patnaik, L. M. (2009). *Soft Computing for Data Mining Applications*. New York, NY: Springer.

Voet, D., & Voet, J. G. (2011a). *Biochemistry* (4th ed.). United States of America: John Wiley & Sons.

Voet, D., & Voet, J. G. (2011b). *Biochemistry* (Fourth edi). United States of America: John Wiley & Sons.

von Lintig, J. (2010). Colors with functions: elucidating the biochemical and molecular basis of carotenoid metabolism. *Annual Review of Nutrition*, *30*, 35–56. http://doi.org/10.1146/annurev-nutr-080508-141027

Westhof, E. (2010). The amazing world of bacterial structured RNAs. *Genome Biology*, *11*(3), 108. http://doi.org/10.1186/gb-2010-11-3-108

Wheeler, T. J., Clements, J., & Finn, R. D. (2014). Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinformatics*, *15*(7), 2–9. http://doi.org/10.1186/1471-2105-15-7

Wieland, B., Feil, C., Gloria-Maercker, E., Thumm, G., Lechner, M., Bravo, J.-M., … Götz, F. (1994). Genetic and Biochemical Analyses of the Biosynthesis of the Yellow Carotenoid 4 , 4 ' -Diaponeurosporene of Staphylococcus aureus. *Journal of Bacteriology*, *176*(24), 7719–7726.

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(11), 5088–5090. http://doi.org/10.1073/pnas.74.11.5088

Yu, Y.-K., & Altschul, S. F. (2005). The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. *Bioinformatics*, *21*(7), 902–911. http://doi.org/10.1093/bioinformatics/bti070

Anexos

Anexo 1. Proteomas seleccionadas para el estudio. Disponible en archivo Anexo1.ods

Anexo 2. Información sobre los perfiles HMM. Se muestran datos sobre los proteomas de prueba; rango de selección (en bit score); número de secuencias alineadas y longitud del perfil HMM. Disponible en el archivo Anexo 2.ods

Anexo 3. Matriz con la abundancia total de las proteínas de la BC en cada uno de los proteomas. Disponible en el archivo Anexo 3.ods

Anexo 4. Matriz con la abundancia de las proteínas de la BC, respecto al total de proteínas en los proteomas. Disponible en Anexo 4.ods

Anexo1. Proteomas seleccionados para el presente estudio. Se muestr
En rojo se muestran las especies en en las que no se encontró ningún

Proteomas analizadas por Ciccarrelli	NCBI Tax.ID
Acidobacterium capsulatum ATCC 51196	240015
Agrobacterium tumefaciens Cereon	181661
Agrobacterium tumefaciens WashU	180835
Aquifex aeolicus	63363
Bacillus anthracis str. Ames	198094
Bacillus cereus ATCC 10987	222523
Bacillus cereus ATCC 14579	226900
Bacillus halodurans	86665
Bacillus subtilis	1423
Bacteroides thetaiotaomicron	818
Bdellovibrio bacteriovorus	959
Bifidobacterium longum	216816
Bordetella bronchiseptica	518
Bordetella parapertussis	519
Bordetella pertussis	520
Borrelia burgdorferi	139
Bradyrhizobium japonicum	375
Brucella melitensis	29459
Brucella melitensis biovar Suis	29461
Buchnera aphidicola (Acyrthosiphon pisun	118099
Buchnera aphidicola (Baizongia pistaciae)	135842
Buchnera aphidicola (Schizaphis graminu	98794
Campylobacter jejuni	197
Candidatus Blochmannia floridanus	203907
Solibacter usitatus Ellin6076	234267
Caulobacter vibrioides	155892
Chlamydia muridarum	83560
Chlamydia trachomatis	813
Chlamydophila caviae	83557
Chlamydophila pneumoniae AR39	115711
Chlamydophila pneumoniae CWL029	115713
Chlamydophila pneumoniae J138	138677
Chlamydophila pneumoniae TW-183	182082
Chlorobaculum tepidum	1097
Chromobacterium violaceum	536
Clostridium acetobutylicum	1488
Clostridium perfringens	1502
Clostridium tetani	1513
Corynebacterium diphtheriae	1717
Corynebacterium efficiens	152794
Corynebacterium glutamicum	1718
Corynebacterium glutamicum ATCC 1303	196627
Coxiella burnetii	777
Dehalococcoides ethenogenes 195	243164

Deinococcus radiodurans	1299
Desulfovibrio vulgaris subsp. vulgaris str.	882
Enterococcus faecalis	1351
Escherichia coli	562
Escherichia coli O157:H7 EDL933	155864
Escherichia coli O157:H7	83334
Escherichia coli O6	217992
Fibrobacter succinogenes subsp. succino	59374
Fusobacterium nucleatum subsp. nucleat	76856
Geobacter sulfurreducens	35554
Gemmata obscuriglobus UQM 2246	214688
Gloeobacter violaceus	33072
Haemophilus ducreyi	730
Haemophilus influenzae	727
Helicobacter hepaticus	32025
Helicobacter pylori	210
Helicobacter pylori J99	85963
Lactobacillus johnsonii	33959
Lactobacillus plantarum	1590
Lactococcus lactis subsp. lactis	1360
Leptospira interrogans serovar Copenhag	44275
Leptospira interrogans	173
Listeria innocua	1642
Listeria monocytogenes	1639
Listeria monocytogenes str. 4b F2365	265669
Mesorhizobium loti	381
Mycobacterium avium subsp. paratubercu	1770
Mycobacterium bovis	1765
Mycobacterium leprae	1769
Mycobacterium tuberculosis CDC1551	83331
Mycobacterium tuberculosis H37Rv	83332
Mycoplasma gallisepticum	2096
Mycoplasma genitalium	2097
Mycoplasma mobile 163K	267748
Mycoplasma mycoides subsp. mycoides	*2102*
Mycoplasma penetrans	28227
Mycoplasma pneumoniae	2104
Mycoplasma pulmonis	2107
Neisseria meningitidis serogroup B	491
Neisseria meningitidis serogroup A	65699
Nitrosomonas europaea	915
Nostoc sp. PCC 7120	103690
Oceanobacillus iheyensis	182710
Onion yellows phytoplasma	100379
Pasteurella multocida	747
Photobacterium profundum	74109
Photorhabdus luminescens subsp. laumor	141679
Pirellula sp.	117
Porphyromonas gingivalis	837

Prochlorococcus marinus	1219
Prochlorococcus marinus subsp. pastoris	59919
Prochlorococcus marinus str. MIT 9313	74547
Pseudomonas aeruginosa	287
Pseudomonas putida KT2440	160488
Pseudomonas syringae pv. tomato	323
Ralstonia solanacearum	305
Rhodopseudomonas palustris	1076
Rickettsia conorii	781
Rickettsia prowazekii	782
Salmonella typhi	601
Salmonella enterica subsp. enterica serov	209261
Salmonella typhimurium	602
Shewanella oneidensis	70863
Shigella flexneri 2a str. 2457T	198215
Shigella flexneri	623
Sinorhizobium meliloti	382
Staphylococcus aureus subsp. aureus Mu	158878
Staphylococcus aureus subsp. aureus M	196620
Staphylococcus aureus subsp. aureus N3	158879
Staphylococcus epidermidis	1282
Streptococcus agalactiae serogroup V	216466
Streptococcus agalactiae serogroup III	216495
Streptococcus mutans	1309
Streptococcus pneumoniae R6	171101
Streptococcus pneumoniae	1313
Streptococcus pyogenes	1314
Streptococcus pyogenes MGAS315	198466
Streptococcus pyogenes MGAS8232	186103
Streptococcus pyogenes SSI-1	193567
Streptomyces avermitilis	33903
Streptomyces coelicolor	1902
Synechococcus elongatus	32046
Synechococcus sp. WH 8102	84588
Synechocystis sp. PCC 6803	1148
Thermoanaerobacter tengcongensis	119072
Thermotoga maritima	2336
Thermus thermophilus HB27	262724
Treponema denticola	158
Treponema pallidum	160
Tropheryma whipplei str. Twist	203267
Tropheryma whipplei TW08/27	218496
Ureaplasma parvum	134821
Vibrio cholerae	666
Vibrio parahaemolyticus	670
Vibrio vulnificus	672
Vibrio vulnificus YJ016	196600
Wigglesworthia glossinidia endosymbiont	36870
Wolbachia sp. wMel	66077

Wolinella succinogenes	844
Xanthomonas axonopodis pv. citri	92829
Xanthomonas campestris pv. campestris	340
Xylella fastidiosa	2371
Xylella fastidiosa Temecula1	183190
Yersinia pestis biovar Medievalis str. 9100	229193
Yersinia pestis	632
Yersinia pestis KIM	187410

ra el código de ensamblado (Assembly) para en los genomas en los que no se descargo el proteoma predicho c
genoma asociado en el servido FTP, del NCBI

Proteomas analizadas en este estudio	NCBI Tax.ID	Reference Genome	Assembly
Acidobacterium capsulatum ATCC 51196 uid59127	240015	1086	
Agrobacterium fabrum C58 uid57865	*176299*	13606	
Agrobacterium_fabrum_C58_uid57865	*176299*	13606	
Aquifex aeolicus VF5 uid57765	*224324*	1049	
Bacillus anthracis Ames Ancestor uid58083	*198094*	181	
Bacillus cereus ATCC 10987 uid57673	*222523*	157	GCA_000008005.1
Bacillus cereus ATCC 14579 uid57975	*226900*	157	GCA_000007825.1
Bacillus halodurans C 125 uid57791	*272558*	1055	
Bacillus subtilis 168 uid57675	*224308*	665	
Bacteroides thetaiotaomicron VPI 5482 uid62913	*226186*	1093	
Bdellovibrio bacteriovorus Tiberius uid182482	*1069642*	1643	
Bifidobacterium longum NCC2705 uid57939	*206672*	183	
Bordetella bronchiseptica 253 uid178913	*568707*	1006	
Bordetella parapertussis Bpp5 uid177516	*1208660*	1007	
Bordetella pertussis Tohama I uid57617	*257313*	1008	
Borrelia burgdorferi B31 uid57581	*224326*	738	
Bradyrhizobium japonicum USDA 6 uid158851	*1037409*	811	
Brucella melitensis bv 1 16M uid57735	*224914*	943	
Brucella suis 1330 uid57927	*204722*	806	
Buchnera aphidicola APS Acyrthosiphon pisum uid	*118099*	170	GCA_000009605.1
Buchnera aphidicola Bp Baizongia pistaciae uid57	*224915*	170	GCA_000007725.1
Buchnera aphidicola str. Sg (Schizaphis graminum)	*198804*	170	GCA_000007365.
Campylobacter jejuni NCTC 11168 ATCC 700819	*192222*	57587	
Candidatus Blochmannia floridanus uid57999	*203907*	1115	
Candidatus Solibacter usitatus Ellin6076 uid58139	*234267*	1167	
Caulobacter crescentus CB15 uid57891	*190650*	941	
Chlamydia muridarum Nigg uid57785	*83560*	1053	
Chlamydia trachomatis D UW 3 CX uid57637	*272561*	471	
Chlamydophila caviae GPIC uid57783	*83557*	1052	
Chlamydophila pneumoniae AR39 uid57809	*115711*	171	GCA_000091085.1
Chlamydophila pneumoniae CWL029 uid57811	*115713*	171	GCA_000008745.1
Chlamydophila pneumoniae J138 uid57829	*138677*	171	GCA_000011165.1
Chlamydophila pneumoniae TW 183 uid57997	*182082*	171	GCA_000007205.1
Chlorobaculum parvum NCIB 8327 uid59185	*517417*	1075	
Chromobacterium violaceum ATCC 12472 uid58001	*243365*	1117	
Clostridium acetobutylicum ATCC 824 uid57677	*272562*	1022	
Clostridium perfringens 13 uid57681	*195102*	158	
Clostridium tetani E88 uid57683	*212717*	1098	
Corynebacterium diphtheriae 31A uid84309	*698962*	1025	
Corynebacterium efficiens YS 314 uid62905	*196164*	1076	
Corynebacterium glutamicum ATCC 13032 uid1937(	*196627*	469	GCA_000011325.1
Corynebacterium glutamicum ATCC 13032 substr. K	*1204414*	469	GCA_000382905.1
Coxiella burnetii RSA 493 uid57631	*227377*	543	
Dehalococcoides_CBDB1_uid58413	*255470*	1048	

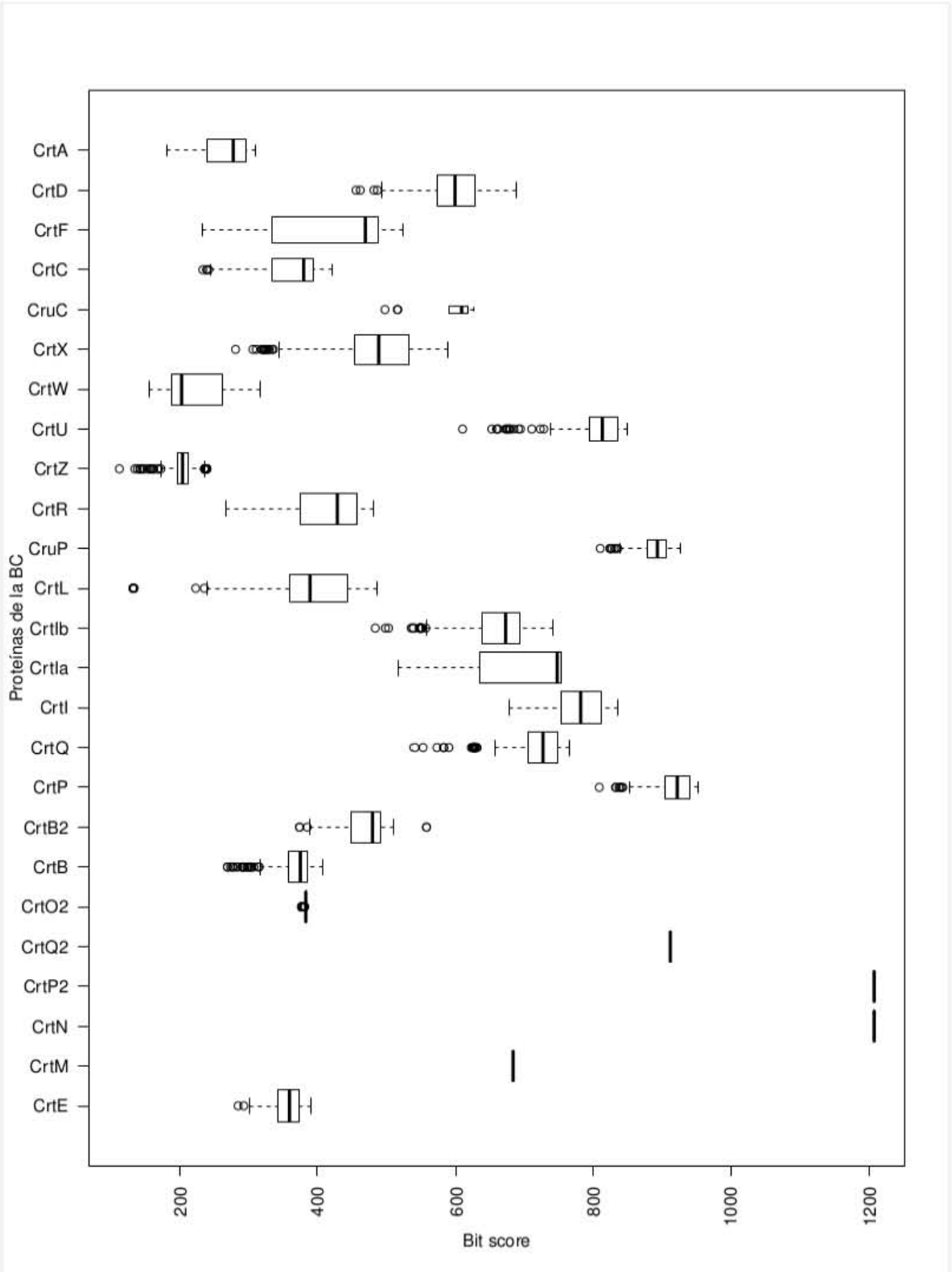
Deinococcus radiodurans R1 uid57665	*243230*	1020	
Desulfovibrio vulgaris Hildenborough uid57645	*882*	654	
Enterococcus faecalis V583 uid57669	*226185*	808	
Escherichia coli K 12 substr MG1655 uid57779	*879462*	167	GCA_000005845.2
Escherichia coli O157 H7 EDL933 uid57831	*155864*	167	GCA_000732965.1
Escherichia coli O157 H7 Sakai uid57781	*386585*	167	GCA_000008865.1
-			
Fibrobacter succinogenes S85 uid41169	*59374*	932	
Fusobacterium nucleatum ATCC 25586 uid57885	*190304*	180	
Geobacter sulfurreducens PCA uid57743	*243231*	1042	
-	*214688*	1623	
Gloeobacter violaceus PCC 7421 uid58011	*251221*	1124	
Haemophilus ducreyi 35000HP uid57625	*233412*	1010	
Haemophilus influenzae Rd KW20 uid57771	*71421*	165	
Helicobacter hepaticus ATCC 51449 uid57737	*235279*	1102	
Helicobacter pylori 26695 uid57787	*85962*	169	GCA_000008525.1
Helicobacter pylori J99 uid57789	*85963*	169	GCA_000008785.1
Lactobacillus johnsonii NCC 533 uid58029	*257314*	1644	
Lactobacillus plantarum WCFS1 uid62911	*220668*	1108	
Lactococcus lactis Il1403 uid57671	*272623*	156	
Leptospira interrogans serovar Copenhageni Fiocru	*267671*	179	GCA_000007685.1
Leptospira interrogans serovar Lai 56601 uid57881	*189518*	179	GCA_000092565.1
Listeria innocua Clip11262 uid61567	*272626*	1024	
Listeria monocytogenes EGD e uid61583	*1334565*	159	GCA_000196035.1
Listeria monocytogenes serotype 4b F2365 uid5768	*265669*	159	GCA_000008285.1
Mesorhizobium loti MAFF303099 uid57601	*266835*	1003	
Mycobacterium avium paratuberculosis K 10 uid576	*262316*	160	
Mycobacterium bovis AF2122 97 uid57695	*233413*	161	
Mycobacterium leprae TN uid57697	*272631*	903	
Mycobacterium tuberculosis CDC1551 uid57775	*83331*	166	GCA_000008585.1
Mycobacterium tuberculosis H37Rv uid170532	*83332*	166	GCA_000195955.2
Mycoplasma gallisepticum R low uid57993	*710127*	1113	
Mycoplasma genitalium G37 uid57707	*243273*	474	
Mycoplasma mobile 163K uid58077	*267748*	1651	
Mycoplasma mycoides SC PG1 uid58031	*272632*	720	
Mycoplasma penetrans HF 2 uid57729	*272633*	1037	
Mycoplasma pneumoniae M129 B7 uid185759	*272634*	1028	
Mycoplasma pulmonis UAB CTIP uid61569	*272635*	1029	
Neisseria_meningitidis_MC58_uid57817	*122586*	172	GCA_000008805.1
Neisseria_meningitidis_Z2491_uid57819	*122587*	172	GCA_000009105.1
Nitrosomonas europaea ATCC 19718 uid57647	*228410*	1013	
Nostoc PCC 7120 uid57803	*103690*	13531	
Oceanobacillus iheyensis HTE831 uid57867	*221109*	1069	
Onion yellows phytoplasma OY M uid58015	*262768*	1125	
Pasteurella multocida Pm70 uid57627	*272843*	912	
Photobacterium profundum SS9 uid62923	*298386*	1193	
Photorhabdus luminescens laumondii TTO1 uid615	*243265*	1123	
Pirellula staleyi DSM 6068 uid43209	*530564*	1760	
Porphyromonas gingivalis W83 uid57641	*242619*	714	

Prochlorococcus marinus CCMP1375 uid57995	*244376*	164	GCA_000007925.1
Prochlorococcus marinus pastoris CCMP1986 uid57	*59919*	164	GCA_000011465.1
Prochlorococcus marinus MIT 9313 uid57773	*74547*	164	GCA_000011485.1
Pseudomonas aeruginosa PAO1 uid57945	*208964*	57945	
Pseudomonas putida KT2440 uid57843	*160488*	174	
Pseudomonas syringae tomato DC3000 uid57967	*223283*	2253	
Ralstonia solanacearum GMI1000 uid57593	*267608*	490	
Rhodopseudomonas palustris CGA009 uid62901	*258594*	508	
Rickettsia conorii Malish 7 uid57633	*272944*	1011	
Rickettsia prowazekii Madrid E uid61565	*272947*	737	
Salmonella_enterica_serovar_Typhi_CT18_uid5779	*90371*	152	GCA_000195995.1
Salmonella_enterica_serovar_Typhi_Ty21a_uid201	*99287*	152	GCA_000385905.1
Salmonella_enterica_serovar_Typhimurium_LT2_ui	*90370*	152	GCA_000006945.1
Shewanella oneidensis MR 1 uid57949	*211586*	1082	
Shigella flexneri 2a 2457T uid57991	*198215*	182	GCA_000007405.1
Shigella flexneri 2a 301 uid62907	*198214*	182	GCA_000006925.2
Sinorhizobium meliloti 1021 uid57603	*266834*	1004	
Staphylococcus aureus Mu50 uid57835	*158878*	154	GCA_000009665.1
Staphylococcus aureus MW2 uid57903	*196620*	154	GCA_000011265.1
Staphylococcus aureus N315 uid57837	*158879*	154	GCA_000009645.1
Staphylococcus epidermidis ATCC 12228 uid57861	*176280*	155	
Streptococcus agalactiae 2603V R uid57943	*208435*	186	GCA_000007265.1
Streptococcus agalactiae NEM316	*216495*	186	GCA_000196055.1
Streptococcus mutans UA159 uid57947	*210007*	856	
Streptococcus pneumoniae R6 uid57859	*171101*	176	GCA_000007045.1
Streptococcus pneumoniae TIGR4	*170187*	176	GCA_000006885.1
Streptococcus pyogenes M1 GAS uid57845	*1314*	175	GCA_000006785.1
Streptococcus pyogenes MGAS315 uid57911	*198466*	175	GCA_000007425.1
Streptococcus pyogenes MGAS8232 uid57871	*186103*	175	GCA_000007285.1
Streptococcus pyogenes SSI 1 uid57895	*193567*	175	GCA_000011285.1
Streptomyces avermitilis MA 4680 uid57739	*227882*	1040	
Streptomyces coelicolor A3 2 uid57801	*100226*	1057	
Synechococcus elongatus PCC 6301 uid58235	*269084*	430	
Synechococcus WH 8102 uid61581	*84588*	13522	
Synechocystis PCC 6803 substr GT I uid157913	*1080228*	13549	
Thermoanaerobacter tengcongensis MB4 uid57813	*273068*	1524	
Thermotoga maritima MSB8 uid57723	*243274*	1034	
Thermus thermophilus HB27 uid58033	*262724*	461	
Treponema denticola ATCC 35405 uid57583	*999430*	1001	
Treponema pallidum Nichols uid57585	*243275*	741	
Tropheryma whipplei str. Twist	*203267*	162	GCA_000007485.1
Tropheryma whipplei TW08 27 uid57961	*218496*	162	GCA_000196075.1
Ureaplasma parvum serovar 3 ATCC 700970 uid577	*273119*	716	
Vibrio cholerae O1 biovar El Tor N16961 uid57623	*686*	505	
Vibrio parahaemolyticus RIMD 2210633 uid57969	*223926*	691	
Vibrio vulnificus CMCP6 uid62909	*216895*	189	GCA_000039765.1
Vibrio vulnificus YJ016 uid58007	*216895*	189	GCA_000009745.1
Wigglesworthia glossinidia endosymbiont of Glossina	*36870*	1066	
Wolbachia endosymbiont of Drosophila melanogaste	*163164*	1065	

Wolinella succinogenes DSM 1740 uid61591	*273121*	1116	
Xanthomonas axonopodis citri 306 uid57889	*1203463*	527	
Xanthomonas campestris ATCC 33913 uid57887	*190485*	57887	
Xylella fastidiosa 9a5c uid57849	*160492*	173	GCA_000006725.1
Xylella fastidiosa Temecula1 uid57869	*183190*	173	GCA_000007245.1
Yersinia pestis biovar Medievalis Harbin 35 uid1585	*229193*	153	GCA_000186725.1
Yersinia pestis CO92 uid57621	*214092*	153	GCA_000009065.1
Yersinia pestis KIM 10 uid57875	*187410*	153	GCA_000006645.1

del genoma de referencia.

Anexo 5. Diagramas de caja que representan los rangos de selección de los perfiles HMM.



Anexo 6. Logos HMM generados a partir del alineamiento múltiple de secuencias.