UNIVERSIDAD DE LA REPUBLICA

FACULTAD DE CIENCIAS

PEDECIBA

Secuenciado y estudios evolutivos del genoma de la bacteria *Delftia* sp. JD2 y de la familia Comamonadaceae

Eugenio Salvador Jara Tellechea

Tesis de Maestría en Genética

URUGUAY

2016

UNIVERSIDAD DE LA REPUBLICA

FACULTAD DE CIENCIAS

PEDECIBA

Secuenciado y estudios evolutivos del genoma de la bacteria *Delftia* sp. JD2 y de la familia Comamonadaceae

Eugenio Salvador Jara Tellechea

Dr. Héctor Musto
Director de la tesis

Dr. Andrés Iriarte
Co-Director de la tesis

PEDECIBA Biología,
Maestría en Genética
Orientador: Héctor Musto
Co-orintador: Andrés Iriarte

INDICE

Agradecimiento.

RESUMEN

La familia Comamonadaceae pertenece a la subdivisión beta del filo Proteobacteria. Los miembros de esta familia tienen una gran versatilidad para degradar compuestos orgánicos e inorgánicos. Hay muy escasos artículos sobre las funciones relacionadas con la evolución molecular de este grupo. En este trabajo se presenta el análisis del patrón de uso de codones en los genomas completamente secuenciados y borradores de las bacterias que pertenecen a esta familia. Describimos e interpretamos en un marco filogenético el efecto de la selección natural en la traducción actuando en los niveles de velocidad y precisión. Hemos encontrado que hay un cierto nivel de variación en la fuerza de la selección entre los microorganismos analizados, lo que probablemente se asocia con la inercia filogenética. El sesgo del uso de codones no se conserva a través de la evolución en la familia de genes altamente expresados, pero sí a nivel de género, lo que sugiere un papel importante de la selección negativa en este nivel. Se identificaron codones óptimos. Se analizaron la estructura secundaria de los ARNm de los genes de las proteínas ribosomales y del resto de los genes del genoma. Se encontró que los mensajeros codificantes de las proteínas ribosomales presentan un valor de energía libre de plegamiento menor en comparación con el resto de los genes. Este patrón se explica como una estrategia en la traducción de los genes de alta expresión, favoreciendo la traducción rápida. Las regiones de las proteínas de membranas presentaron un enriquecimiento de codones terminados en G. Este patrón es conservado en forma general por todos los organismos en estudio.

INTRODUCCIÓN GENERAL

1.1 Estructura del genoma procariota

El genoma de los procariotas se caracteriza por poseer una composición homogénea (contenido en GC) y gran densidad de genes, teniendo muy pocas secuencias intergénicas. La replicación del ADN no es un mecanismo simétrico, sino que cada hebra se replica en forma distinta exponiéndose a diferentes condiciones que generan una tasa de mutación diferencial entre ellas (Frederico et al. 1990). Esto causa un desbalance en las proporciones de las bases y de genes entre las hebras (Francino y Ochman 2001; Frederico et al. 1990). La replicación de la hebra rezagada se genera por fragmentos de Okazaki quedando temporalmente como una hebra de ADN simple, que va a tender a acumular una mayor cantidad de mutaciones respecto a la hebra líder (Frederico et al. 1990). La replicación de la hebra líder se genera de forma continua por lo que nunca está como hebra simple, sufriendo de esta manera una menor cantidad de mutaciones con respecto a la hebra rezagada (Frederico et al. 1990). Sin embargo, se ha propuesto que gracias a la especificidad que posee la ADN ligasa al unir los fragmentos de Okazaki, puede disminuir la probabilidad de acumular mutaciones y conferir un grado de edición de los errores (Housby y Southern 1998). De esta forma se plantea que la hebra rezagada va a tender acumular menos cambios de bases con respecto a la hebra líder. La hipótesis más aceptada para poder explicar la asimetría de bases entre las hebras es debido al sesgo mutacional y en especial, a la mutación de tipo desaminación de citosina (Frank y Lobry 1999). Generalmente la hebra líder tiene más contenido de Guanina (G) y Timina (T) respecto Adenina (A) y Citosina (C) y de genes con respecto a la hebra rezagada; a estos fenómenos se los llama “GC-skew” (Lobry 1996) y “Gen-skew”, respectivamente. Estas características en la organización del genoma procariota son explicadas como una consecuencia en relación a la maquinaria de replicación y transcripción, siendo favorecida aquella configuración en donde ambos procesos son codireccionales. La diferencia en las bases se explicaría por una diferencia en la tasa mutacional de ambas hebras. Dicho de otra forma, si la ARN polimerasa transcribe al mismo tiempo un gen situado en la hebra rezagada va a presentar una mayor probabilidad de chocar con la ADN polimerasa que replica el ADN en el sentido contrario generando transcriptos abortivos (Rocha 2004b). Si la ARN polimerasa transcribe en la misma dirección que la ADN polimerasa la probabilidad de generar un transcripto abortivo disminuye y la replicación es más rápida (Rocha 2004b), siendo selectivamente favorecidos los genes situados en la cadena líder.

El aumento en la disponibilidad de genomas secuenciados ha permitido aproximarnos a comprender las dinámicas de las fuerzas evolutivas que actúan en éstos. El incremento en el número de genes en los genomas dependería principalmente de la transferencia horizontal y no tanto de la duplicación génica (Treangen y Rocha 2011). Existe una relativa uniformidad en el cromosoma en términos de densidad génica, composición de secuencia, etc. Los genes son en muchos casos transcriptos en conjunto (operones) y los procesos celulares están altamente asociados dada la inexistencia (en general) de compartimentos subcelulares definidos. El tamaño del genoma se correlaciona con el número de genes (Mira et al. 2001), y a la frecuencia de transferencia horizontal de genes (Cordero y Hogeweg 2009), con la aerobiosis y el contenido de GC genómico (Naya et al. 2002). Sin embargo, existe probablemente una correlación negativa entre el tamaño del genoma y la estructura y la conservación de los operones (Nuñez et al. 2013). En conjunto, estas características hacen que los cromosomas de los procariotas tengan, al mismo tiempo, una organización compleja y conservada y un repertorio génico altamente plástico (Rocha 2008). En definitiva, es altamente aceptado que la organización de la información genética es un carácter bajo selección, en donde la mayoría de los reordenamientos llevan a una disminución en la adaptabilidad de los organismos que lo poseen. La plasticidad y la organización del cromosoma son en última instancia moldeados por la selección, la deriva y la tasa mutacional, dependiendo de procesos celulares y ecológicos. La comprensión de estas dinámicas es posible mediante estudios de genómica comparativa, siempre y cuando estén disponibles un número significativo de organismos, idealmente siendo representativos de cierta variabilidad ecológica, en un marco filogenético robusto (Koonin y Wolf 2008).

1.2 Uso de codones sinónimos

En el código genético "universal", todos los aminoácidos son codificados por más de un codón a excepción de los aminoácidos Met y Trp que son codificados por un solo triplete. Los codones que codifican a un mismo aminoácido son llamados codones sinónimos. Este comportamiento del código genético en el que hay más de un codón para la codificación de un aminoácido se le conoce como redundancia en el código. Este presenta una gran conservación entre los organismos, sin embargo se han reportados algunas excepciones (Santos et al. 2004). En un principio se pensaba que los cambios sinónimos no tenían ningún efecto o eran neutrales en la síntesis proteica, porque mantenían la codificación del aminoácido, por lo que se tomó como uno de los pilares de la evolución neutral (King y Jukes 1969). Sin embargo, no son utilizados en una misma frecuencia, ya que en general se tiende a preferir al codón sinónimo que se corresponda de mejor manera con el

ARNt más abundante (Gouy y Gautier 1982; Ikemura 1985; Kanaya et al. 1999). Granthan y colaboradores realizaron análisis de correspondencia a las secuencias de ARNm de diferentes organismos, y observaron agrupamientos por el tipo de genoma. De esta manera se postuló la "hipótesis genómica", en la que cada tipo de organismo posee un sesgo particular en el uso de codones sinónimos (Grantham et al. 1980). Al ir aumentando la cantidad de secuencias disponibles, se ha demostrado que no sólo existen sesgos en el uso de codones a nivel inter genómico sino que también existen a nivel intra genómico. Es decir, los codones sinónimos no se utilizan en una misma frecuencia al compararseles entre organismos y entre genes de la misma especie (Gouy y Gautier 1982; Ikemura 1985; Sharp et al. 2005). Se ha establecido que existe en ciertos genes un sesgo hacia codones llamados óptimos, estando asociado este fenómeno al nivel de expresión (Gouy y Gautier 1982). Por lo tanto, los genes más expresados (los que codifican por ejemplo para proteínas ribosomales, factores de elongación) van a presentar una sesgo mayor en la preferencia de determinados codones sinónimos, fenómeno generado y mantenido por la selección (Goetz y Fuglsang 2005). Se ha demostrado que la elección de los codones preferidos por la selección son particulares en cada organismo (Sharp et al. 2005; Sharp y Li 1986). Qué factores generan que existan tales diferencias en la preferencia de ciertos codones sinónimos en los organismos, es todavía un tema sin resolver. Se ha encontrado una correlación entre el contenido de GC en las regiones intergénicas y los codones preferidos. En otras palabras, los organismos que presentan regiones intergénicas ricas en GC, van a presentar codones preferidos ricos en GC, aunque este patrón se debilita en los genomas pobres en GC (Hershberg y Petrov 2009). A nivel intra génico, se ha demostrado que existe una preferencia para ciertos codones sinónimos sobre otros (Cortazzo et al. 2002; Thanaraj y Argos 1996; Tuller et al. 2010), lo que podría afectar el plegamiento de las proteínas produciendo cambios en su actividad biológica y solubilidad (Cortazzo et al. 2002; Marin 2008). De esta manera el sesgo en el Uso de Codones Sinónimos (UCS) puede ser estudiado a nivel inter e intra genómico e intra génico.

1.2.1 Fuerzas que modelan el uso de codones sinónimos (UCS)

En términos generales, existen dos visiones que intentan explicar el origen y la dinámica en el comportamiento del sesgo en el UCS: la seleccionista y la neutralista.

La visión seleccionista propone que el sesgo en el uso codones sinónimos contribuye a la eficacia y fidelidad en la traducción proteica, siendo éste el resultado de un balance entre el sesgo mutacional, la deriva génica y la presión selectiva en una población finita (Bulmer 1991). La teoría de selección-mutación-deriva, explica la permanencia de los codones sinónimos no óptimos como

resultado de un equilibrio dinámico entre la selección, la deriva génica y la mutación. La selección actúa a favor de un codón óptimo para cada aminoácido mientras que la mutación y la deriva génica permiten la persistencia de los codones no óptimos. Se ha demostrado que los genes que codifican las proteínas ribosomales y otros genes de alto nivel de expresión prefieren un subgrupo de codones sinónimos a los que se les considera como codones óptimos (Gouy y Gautier 1982). La preferencia de los codones sinónimos óptimos generaría una tasa de elongación más rápida y con alta fidelidad (Bulmer 1991; Hershberg y Petrov 2008). Habitualmente los genes más expresados son los que presentan un sesgo más intenso en el UCS, siendo el mismo establecido y mantenido por la selección. Es de destacar que estas ideas generales han sido reafirmadas en diferentes formas (Akashi 1994; Musto et al. 2003). Por lo tanto, el uso de ciertos codones sinónimos pueden estar implicado en la velocidad de traducción (Tuller et al. 2010) y en la precisión de la misma (Akashi 1994).

Por otra parte, el neutralismo explica el UCS como consecuencia del sesgo mutacional de cada organismo y la mutabilidad de los codones, siendo la deriva la responsable en general de fijar las variaciones (Kimura 1968). Las mutaciones silenciosas serían invisibles a la selección natural ya que no modifican el aminoácido codificado.

El principal factor que determina el uso mayoritario de determinados codones sinónimos a nivel global en el genoma es el sesgo mutacional (Chen et al. 2004; Knight et al. 2001), y este sesgo en el UCS puede ser explicado en cada genoma a partir del contenido de GC en su secuencias intergénicas (Chen et al. 2004). Los procariotas presentan un gran rango de GC en la secuencias del genoma variando aproximadamente entre un 25% y 75%, principalmente afectando la tercera posición de los codones (Muto y Osawa 1987); por lo que, la variación inter-génomica en el uso de codones sinónimos, se puede explicar entonces fácilmente por un proceso mutacional genómico global más que selectivo. Por otra lado, la visión seleccionista, propone que el porcentaje de GC de un organismo va a estar influenciado por el sesgo mutacional pero también por la selección natural (Musto et al. 2006; Naya et al. 2002). Se puede establecer, por lo tanto, un equilibrio dinámico entre el sesgo mutacional y la selección natural, el cual determinaría el porcentaje de GC que posee un genoma y por lo tanto, el sesgo en el UCS.

Por otro lado, existen algunas claras indicaciones de que la selección natural debe estar también involucrada en el sesgo en el UCS a nivel intragénico. Por ejemplo, la presión mutacional no puede explicar por qué en muchos casos los codones más frecuentemente utilizados por los genes de alta expresión, son aquellos que son reconocidos por el ARN de transferencia (ARNt) isoaceptor más

abundante (Ikemura 1985; Kanaya et al. 2001; Kanaya et al. 1999). Los ARNt sufren modificaciones postranscripcionales, afectando la elección de cierto codón sinónimo por otro (Gustilo et al. 2008; Santos et al. 2004). Aquellas bacterias que poseen un tiempo generacional corto tienen más cantidad pero menos diversidad de genes de ARNt que las bacterias de generación más largas (Rocha 2004a). Esto condiciona el UCS, optando por el codón que minimice el tiempo requerido para que el ARNt isoaceptor llegue al sitio A del ribosoma. En general van a ser favorecidas las interacciones del tipo Watson-Crick (WC) entre la tercera base del codón y la primera base del anticodón (Ikemura 1981), lo que aumenta la especificidad de la interacción de las bases del codón con el ARNt. Varios modelos intentan explicar el uso diferencial de codones sinónimos respecto al ARNt isoaceptor (Grosjean and Fiers 1982; Rocha 2004a). La interacción del ARNt isoaceptor con el codón en el sitio A del ribosoma es mediante difusión. Se ha observado que una estrategia para poder maximizar en tiempo y fidelidad es la reutilización del ARNt, ya que el ARNt luego de ser utilizado queda en la cercanía del ribosoma, aumentando la probabilidad de ser, una vez cargado, usado nuevamente (Cannarozzi et al. 2010). De acuerdo con esto, los genes que utilizan los codones reconocidos por los ARNt isoaceptores más abundantes, serán traducidos más eficientemente y con menos errores. Sin embargo, se ha demostrado que cambios en la concentración ya sea por mutaciones o pérdida de genes de ARNt no dieron lugar a cambios significativos en la eficiencia de la elongación (Pop et al. 2014). De esta manera no queda claro cuáles son los beneficios a nivel de la eficacia en la traducción y elongación, y la implicancia del sesgo en uso de codones sinónimos y la concentración de los ARNt.

El ARNm no solo lleva consigo el código que codifica a los aminoácidos de una proteína, sino que presenta varias particularidades que pueden afectar la síntesis de la proteína, por ejemplo, la formación de determinada estructuras secundarias a nivel del ARNm (“stems” o “loops”). Se ha encontrado que en el extremo 5’ los mensajeros presentan una menor estructura secundaria, la que está vinculada con la tasa de iniciación (Pop et al. 2014). La mínima energía libre (MFE) de una molécula de ARN es afectada por el número, composición y arreglo de los nucleótidos en la secuencia (Trotta 2014). Seffens y Digby (1999) demostraron que la MFE está relacionada con la longitud del ARN: cuanto más larga sea la secuencia del ARN, más negativa va a ser la MFE. Las secuencias de ARNm nativos presentan un valor de MFE más negativo en forma general con respecto a las secuencias de ARNm hechas en forma aleatoria. Este patrón podría estar reflejado en el UCS, favoreciendo las estructuras que contribuyen con la estabilidad en el plegamiento del ARNm (Seffens y Digby 1999). Se ha demostrado que existe una conservación en la estructura secundaria de los ARNm, por lo que las mutaciones que alteren a la estabilidad del ARNm son eliminadas por la selección (Chursov et al. 2013). Estructuras estrechamente plegadas en el ARNm

llevan a una dificultad en la iniciación de la traducción y por lo tanto se disminuye la síntesis proteica (Kudla et al. 2009; Pop et al. 2014). Sin embargo se ha demostrado una correlación positiva entre el plegamiento del ARNm con las proteínas más expresadas (Gorochowski et al. 2015; Park et al. 2013). Además, recientemente se ha encontrado una posible asociación entre las regiones del ARNm que contienen elementos de estructura secundaria estable con regiones de proteínas que codifican dominios compactos y proteínas de gran tamaño (Faure et al. 2016).

Por lo tanto resulta razonable concluir que el ARNm no sólo contiene la información de los aminoácidos a codificar sino que posee capas de información que pueden influir en la síntesis proteica. Se ha observado que ciertos grupos de proteínas tienen un uso particular de codones sinónimos o presentan un sesgo en ciertas regiones (Thanaraj y Argos 1996; Tuller et al. 2010). Las proteínas de membrana serían un caso particular e interesante por sus funciones en la célula.

1.3 Proteínas de membrana

Las proteínas de membrana están constituidas por regiones hidrofóbicas (regiones transmembrana, de aquí en adelante TM) y otras que las separan (no transmembrana, NOTM). Estas proteínas son ricas en los aminoácidos hidrofóbicos Leu, Met, Cys, Val, Phe, Ile (Hessa et al. 2005), siendo utilizados con mayor frecuencia en las regiones TM que en las regiones NOTM (Martinez-Gil et al. 2011). Estas se expresan en general en menor cantidad que las proteínas "citoplasmáticas", y se estima que las regiones que codifican proteínas de membrana están entre el 20-30% de los marcos de lectura abierta y muestran una relación positiva con el aumento del tamaño del genoma (Boyd et al. 1998). Los codones que codifican a los aminoácidos hidrofóbicos (Hessa et al. 2005) tienen un enriquecimiento en U que llega al 50% de sus bases (Prilusky y Bibi 2009). Se ha demostrado además que las proteínas de membrana son ricas en los aminoácidos hidrofílicos Ser y Tyr (Prilusky y Bibi 2009). La región hidrofóbica está constituida por un conjunto de alfa hélices que atraviesan la membrana y se unen entre sí a través de puentes de hidrógeno con un motivo GXXXG (G es el aminoácido Gly y X cualquier aminoácido hidrofóbico) (Curran y Engelman 2003; Melnyk et al. 2004). Si bien aminoácidos polares como la Thr y Ser no son favorables energéticamente a la interacción entre la membrana plasmática y la proteína de membrana, sirven para poder generar puentes de hidrógeno entre las alfa hélices estabilizando a las mismas en estas proteínas (Dawson et al. 2002). Por otra parte, mutaciones de aminoácidos no polares a polares en las regiones TM generan proteínas aberrantes (Curran y Engelman 2003; Gaucher et al. 2006). La sumatoria de residuos no polares neutralizan los efectos desfavorables de los residuos polares (Hessa et al. 2005).

Las regiones TM están conectadas por giros que pueden ser grandes o pequeños. Las regiones citoplasmáticas tienen un enriquecimiento de aminoácidos con carga negativa, mientras que las regiones extracelulares muestran un enriquecimiento de aminoácidos con carga positiva (Wallin y von Heijne 1998). Se ha encontrado que los aminoácidos más conservados en la regiones TM son Gly, Pro y Tyr (Liu et al. 2002), quizás por no ser hidrofóbicos es que presentan tal comportamiento.

Varios estudios indican que cambios sinónimos pueden afectar la estabilidad del ARNm, la estructura de iniciación del ARNm y el plegamiento de la proteína (Fredrick e Ibba 2010; Marin 2008). Por otra parte se ha sugerido que la síntesis proteica no se da a una velocidad constante (Varenne et al. 1984): recientemente, se ha demostrado con experimentos de perfiles de ribosomas, que la velocidad de traducción no es uniforme en el mismo ARNm o entre distintos ARNm (Ingolia 2016; Ingolia et al. 2009). Existe evidencia que sugiere fuertemente que las proteínas de membrana son sensibles a la utilización de codones sinónimos, afectando la estabilidad del ARNm, viéndose disminuida la expresión de la proteína (Duan et al. 2003; Makino et al. 1997). El plegamiento y la estructura de las proteínas de membrana son también sensibles a la utilización de codones sinónimos (Kimchi-Sarfaty et al. 2007); por ejemplo las alfa hélices son traducidas más rápido con respecto a las hoja beta o los loops (Thanaraj y Argos 1996). En general, se ha encontrado que los primeros 30-50 codones son "raros" (no óptimos") siendo traducidos por ARNt poco abundantes, de esta manera la región 5′ se traduce más lenta que la región 3′, esto tiene como consecuencia que disminuya la probabilidad de estancamiento de los ribosomas en el transcurso de la lectura del ARNm (Tuller et al. 2010). Así, una lectura lenta en las región 5' genera una menor frecuencia de ribosomas en las región 3' y evita posibles colisiones que lleven a la pérdida del ARNm.

El estudio en el UCS se ha hecho extensivo a varios grupos de bacterias con genomas secuenciados. En términos generales se acepta que la selección es operativa en muchos organismos (Iriarte et al. 2013; Sharp et al. 2005), incluso en parásitos (Iriarte et al. 2011). Sin embargo hay familias enteras de organismos que no han sido estudiadas en profundidad. Este es el caso de la familia *Comamonadaceae*, lo que es doblemente extraño dada la importancia o potencial biotecnológico que presentan los organismos de esta familia.

1.4 Características y evolución de la familia Comamonadaceae

Las bacterias de la familia *Comamonadaceae* son Gram-, son parte de las beta proteobacterias y presentan un rango de GC de 52 a 70% (Willems 2014). Las especies pertenecientes a esta familia pueden ser encontradas en diversos hábitats, que van desde el suelo, aguas subterráneas, lodo, agua en procesos industriales (*Albidiferax, Alicycliphilus, Caenimonas, Curvibacter, Comamonas, Delftia, Diaphorobacter, Extensimonas, Giesbergeria, Simplicispira, Hydrogenophaga, Hylemonella, Lampropedia, Limnohabitans, Macromonas, Malikia, Ottowia, Polaromonas, Pseudacidovorax, Pseudorhodoferax, Rhodoferax, Variovorax, Xenophilus*). Algunas son patógenas de plantas (*Acidovorax, Xylophilus*), mientras que otras se han encontrado en muestras clínicas (*Acidovorax, Comamonas, Delftia acidovorans*). Se detectaron bacterias simbióticas del género *Verminephrobacter* en los nefridios (Lund et al. 2010) y en hábitats muy fríos polares o de otro tipo (*Polaromonas*, *Rhodoferax antarcticus*), incluyendo el Océano Antártico (*Polaromonas vacuolata*), y en el suelo desierto (*Ramlibacter*). Son quimioorganotróficas o quimiolitotrofas facultativas con oxidación de H2 o CO. Poseen respiración aeróbica, utilizando por lo tanto en forma general al oxígeno como último aceptor de electrones, con algunas excepciones, donde utilizan nitrato como último aceptor. La temperatura óptima de crecimiento se encuentra en el rango de 36–46 grados Celsius a excepción de las *Polaromonas*, que presentan un rango de temperatura de 4 a 18 grados Celsius (por una descripción completa de la familia, ver Willems 2014). En el momento en que se escribe este trabajo, los genomas completamente secuenciados pertenecientes a este grupo son *Comamona, Acidovorax, Alicycliphilus, Delftia*, *Polaromonas*, *Ramlibacter*, *Rhodoferax* y *Variovorax*.

Las bacterias del género *Comamonas* crecen bien con la mayoría de los ácidos orgánicos y aminoácidos pero utilizan muy pocos azúcares (De Vos et al. 1985). Son capases de degradar una amplia variedad de compuestos complejos aromáticos, esteroides y muchas moléculas orgánicas complejas hechas por el hombre (Willems 2014). Utilizan el oxígeno como aceptor final de electrones. Ha sido descrito que cepas de *C. testosteroni* son capaces de degradar hidrocarburos aromáticos policíclicos, tales como fenantreno, naftaleno y antraceno (Goyal y Zylstra 1996). La cepa *C. testosteroni* es resistente al cadmio y se ha aislado de suelos contaminados con metales pesados (Kanazawa y Mori 1996), y las cepas de *Comamonas* resistentes al níquel, se han aislado de suelos contaminados con níquel-percolado (Stoppel y Schlegel 1995). Debido a su capacidad para degradar una amplia variedad de compuestos orgánicos complejos, las cepas de *Comamonas* son de interés potencial en la biorremediación.

A partir de estudios con el rRNA 16S se identificaron a los miembros del género *Comamonas*, entre otros integrantes de la familia *Comamonadaceae* (Amann et al. 1996). El género se validó en 1985, cuando el nombre fue restablecido (De Vos et al. 1985).

El género *Acidovorax* presenta flagelos y un tipo de respiración aerobia, y su contenido de GC varía entre 62% y 70%. La bacterias de este género constituyen uno de los organismos dominantes del suelo, compost viejo y fuentes de agua dulce. Son capaces de degradar a los compuestos poli-3-hidroxibutirato (PHB) y poli-3-hidroxibutirato-co-3-hidroxivalerato (PHBV) *in vitro* (Mergaert y Swings 1996). *A. avenae subsp. avenae* es capaz de despolimerizar los polímeros termoplásticos sintéticos, tales como poli-butileno succinato-co-butileno adipato y poli-caprolactona y materiales compuestos de poli-caprolactona-almidón, y se ha aplicado con éxito en un sistema *in vitro* en un ensayo de biodegradabilidad (SCANDOLA* et al. 1998). Tres especies de *Acidovorax* son los agentes causales de enfermedades en varias plantas. *A. avenae subsp. avenae* causa síntomas de tizón foliar en muchos miembros de la familia *Poaceae*, incluyendo el maíz (*Zea mays*), caña de azúcar (*Saccharum officinarum*) (Claflin et al. 1989).

Se han comparado las especies del género *Verminephrobacter* que viven como endosimbiontes en las lombrices de tierra, con las de dos cepas de vida libre estrechamente relacionados, *Acidovorax avenae* subsp. *citrulli* AAC00-1 y *Acidovorax* sp. JS42. La comparación reveló que el tamaño del genoma no se redujo y no mostró sesgo hacia A-T en los simbiontes (Kjeldsen et al. 2012).
El género *Rhodoferax* presenta varias formas de obtener la energía para su crecimiento, a partir de la fotosíntesis, respiración aeróbica, o fermentación. Se pueden aislar en agua dulce rica en materia orgánica y presenta flagelos. Este fue incluido dentro de las betaproteobacterias, gracias a un estudio filogenético basado en secuencias de ARN 16S (Hiraishi 1994).
El género *Variovorax* presenta flagelos, una respiración aerobia estricta y puede ser aislado de suelo contaminado. Se han reportado cepas de agua de río que degradan compuestos alifáticos como carbonato de policarbonatos de polihexametileno y carbonato de tetrametileno (Suyama et al. 1998).

El género *Delftia* se creó a partir de análisis filogenéticos y fenotípicos de la bacteria *Comamonas acidovorans*, re-nombrándola *Delftia acidovorans* (Wen et al. 1999). Es un organismo aerobio estricto (Willems 2014), con un contenido de GC que varía entre 67% y 69%. Se ha demostrado que las cepas *Delftia sp.* JD2 y *D. acidovorans* SPH-1 pueden reducir el Cr (VI) a Cr (III) (Garavaglia et al. 2010; Morel et al. 2011), siendo este último menos perjudicial para los organismos (Cheung y Gu 2007). Este género posee una gran versatilidad a la hora de degradar compuestos orgánicos e inorgánicos (Juarez-Jimenez et al. 2010). Es capaz de fijar nitrógeno en vida libre, con una

nitrogenasa que funciona con vanadio (Morel et al. 2011). La contaminación por metales pesados de los suelos a menudo se asocia con deficiencia de hierro en una gama de diferentes especies de plantas (Mishra y Kar 1974). Se ha encontrado en *Delftia sp.* JD2, la producción de sideróforos y la fitohormona ácido indol acético (Morel et al. 2011). Una vez unido el hierro con el sideróforo bacteriano, este complejo puede ser tomado por las plantas y por lo tanto sirven como una fuente de hierro para ellas (Bar-Ness et al. 1992). Las bacterias que promueven el crecimiento de las plantas pueden influir positivamente en el crecimiento y su desarrollo (Castro-Sowinski et al. 2007), de esta manera se podrían sustituir métodos químicos por este tipo de metodologías biológicas, las que son más baratas y menos contaminantes.

Delftia sp JD2 y los organismos de la familia *Comamonadaceae* presentan características potenciales, como para la biorremediación de metales pesados o como inoculantes para leguminosas, haciéndolas atractivas para su estudio evolutivo.

A partir de estos antecedentes, nos propusimos secuenciar y profundizar en el estudio de la estructura, composición exhaustiva genómica y hacer una clasificación filogenética de *Delftia sp.* JD2. Se trabajó con las secuencias de esta bacteria y con los organismos secuenciados completamente y los incompletos ("borradores") pertenecientes de la familia *Comamonadaceae*. A partir del análisis de la secuenciación se estudió cuál era la hebra líder, la rezagada, el orden de los genes comparándolo con el resto de los organismos filogenéticamente cercanos, asociándola con la filogenia de la familia *Comamonadaceae*. Hasta el momento no se ha encontrados trabajos que analicen el uso de codones sinónimos en estos organismos. Como se mencionó anteriormente los organismos de esta familia presentan particularidades interesantes en lo que respecta a la biotecnología o biorremediación. A nivel génico analizamos la hipótesis de posible selección a nivel de traducción en los organismos en estudio, identificando cuáles son los codones óptimos, y por último se cuantificó la tasa evolutiva de los genes ortólogos.

OBJETIVOS

2.1 Objetivo General

Secuenciar y analizar el genoma completo de *Delftia sp* JD2. Comparación de los resultados obtenidos con los ya publicados en otros organismos de la familia *Comamonadaceae*.

2.2 Objetivos específicos.

En este sentido se dará especial énfasis a:

- Ensamblar el genoma de la bacteria *Delftia JD2*, anotar los genes, identificar los genes ortólogos y realizar la reconstrucción filogenética utilizando las proteínas ribosomales.
- Realizar la búsqueda de selección en la traducción en los organismos de la familia *Comamonadaceae*.
- Determinar los codones traduccionalmente óptimos en el organismo y en la familia, en un marco filogenético.
- Determinar si las proteínas de membrana presentan un uso particular de codones sinónimos entre las regiones transmembrana con respecto a las no transmembrana.
- Relacionar la energía mínima libres de plegamiento del ARNm entre los genes que codifican a las proteínas ribosomales con respecto el resto de los genes del genoma.

MATERIALES Y MÉTODOS

3.1 Secuencias y Organismos

Los genomas completamente secuenciados fueron obtenidos vía ftp de la base de datos del NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria). Se utilizaron un total de 18 organismos perteneciente a la familia *Comamonadaceae* completamente secuenciados, 13 "Draft" y la cepa *Delftia sp* JD2.

3.2 Secuenciado y Ensamblado

El ADN genómico se secuenció en Macrogen, Inc., Corea del Sur, utilizando un sistema de Illumina HiSeq 2000. A partir de los resultados de la secuenciación se analizaron los "reads", usando el programa fastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/). Se eliminaron los adaptadores illumina y los "reads" de baja calidad utilizando los programas scythe y sickle, respectivamente; se analizó nuevamente a los "reads" restantes mediante la utilización del programa fastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/). El ensamblado se realizó utilizando el programa Spade 3.0.0 (Bankevich et al. 2012) y luego se evaluó el ensamblado mediante el programa ABACAS (http://sourceforge.net/projects/abacas/files/). Por último se utilizó el genoma de *Delftia acidovorans* SPH-1 como referencia para verificar si el ensamblado es correcto. Se identificaron a los marcos abiertos de lectura (ORF) mediante la utilización del programa online RAST (http://rast.nmpdr.org/) (Aziz et al. 2008). Los posibles genes encontrados a través del programa RAST, fueron validados a partir de un análisis en Blastp utilizando las secuencias aminoacídicas de éstos (Pruitt et al. 2005). Nos quedamos con los Hit que mostraron un e-value < 1x10-20.

3.3 Análisis de Delftia sp. DJ2

Mediante un script realizado en el laboratorio se identificaron los genes ortólogos entre los genomas de las bacterias *Delftia sp.* JD2, *Delftia Cs* 1-4, *Delftia acidovoras* SPH-1, siguiendo el criterio de "Best Reciprocal Hit" entre las proteínas, utilizando la herramienta BLASTp (Altschul et al. 1990).

Luego se identificaron a los genes que se encuentran únicamente en *Delftia sp*. JD2, con respecto a *Delftia Cs* 1-4 y *Delftia acidovorans* SPH-1, utilizando el programa BLASTp, mediante las secuencias codificantes de *Delftia sp.* JD2. Para descartar posibles genes mal anotados, se utilizó la herramienta tBLASTx con los genomas de *Delftia Cs* 1-4 y *Delftia acidovorans* SPH-1.

De los genes únicos que presenta *Delftia sp.* JD2 con respecto a las otras dos *Delftia* completamente secuenciadas se eligió al gen dss.5102, que está vinculado a la resistencia a metales pesados. A partir de este gen se realizó a través de la utilización de la herramienta BLASTp una búsqueda en todos los genomas de Genbank (Marzo 2016). Se realizó una filogenia con los genes que mejor "pegaron" con un e-value -10 con respecto al gen dss.5102, mediante el método máxima verosimilitud, utilizando el programa PHYML 3.0 (Guindon et al. 2010)

Se analizó la similaridad mediante el promedio de la identidad nucleotídica (ANI) entre *Delftia Cs* 1-4 y *Delftia acidovorans* SPH-1. Este índice se utiliza para poder delimitar especies a partir de secuencias nucleotídicas (Goris et al. 2007). Genomas que presenten un valor superior del 95% de ANI, se las puede considerar de la misma especie.

Las predicciones de los operones se obtuvieron de la página the Prokaryotic Operon DataBase (ProOpDB; http://operons.ibt.unam.mx/OperonPredictor/) (Taboada et al. 2012).

3.4 Reconstrucción filogenética

Mediante un script realizado en el laboratorio se identificaron los genes ortólogos entre los genomas de las bacterias de la familia *Comamonadaceae*, siguiendo el criterio de "Best Reciprocal Hit" entre las proteínas, utilizando la herramienta BLASTp (Altschul et al. 1990). Nos quedamos con los genes que presentan un valor de probabilidad de e-10. De esta forma la búsqueda de los genes ortólogos entre las bacterias será muy restrictiva disminuyendo la probabilidad de quedarnos con genes parálogos, o genes que fueron incluidos en el genoma a través de transferencia horizontal. Las proteínas ortólogas fueron alineadas independientemente utilizando el programa ClustalW (Thompson et al. 1994).

A partir de los genes ortólogos se identificaron a las proteínas ribosomales ortólogas tomando en cuenta la notación de cada genoma. Por otra parte, se realizó un blastp entre una base de proteínas ribosomales y las proteínas ribosomales ortólogas, de esta manera nos aseguramos de que estamos trabajando efectivamente con este grupo de proteínas. Finalmente, obtuvimos un total de 50 proteínas ribosomales ortólogas. Para poder realizar la reconstrucción filogenética, se concatenaron los alineamientos de estas proteínas. Se estimó el modelo de sustitución de las secuencias aminoacídicas utilizando el programa Modelgenerator (Keane et al. 2006). El modelo de sustitución aminoacídica que mejor se ajusta a las secuencias de las proteínas ribosomales es el LG+G. Por

último, la reconstrucción filogenética se realizó mediante el método de máxima verosimilitud, utilizando el programa PHYML 3.0 (Guindon et al. 2010).

3.5 Uso de codones sinónimos, análisis básicos del genoma y predicción de ARNt

El uso de codones sinónimos (UCS) fue calculado utilizando el programa CodonW 1.4.2 escrito por Peden y disponible en http://sourceforge.net/projects/codonw/. Los codones que codifican los aminoácidos Met y Trp y los tres "stop" fueron excluidos. Con estos datos se realizó el análisis de correspondencia dentro del grupos (WCA) (Charif et al. 2005; Suzuki et al. 2008), utilizando el paquete ADE4 de R (Thioulouse et al. 1997). Además, para cada gen, se calculó el valor MELP (Supek y Vlahoviček 2005) utilizando INCA2.1 (Supek y Vlahoviček 2004). MELP es un predictor del nivel de expresión, independientemente de la longitud y composición de bases del gen. Este índice mide la similitud de uso de codones entre los genes y un grupo de genes altamente expresados seleccionados (GAE). En todos los casos, se establece la referencia para el cálculo de este índice sólo a los genes que codifican a las proteínas ribosomales.

Se utilizó el estadístico χ^2 de contingencia para determinar los codones más utilizados (codones óptimos).

El coeficiente de selección (S) en el uso de codones fue calculado de acuerdo a Sharp et al. (2005), utilizando como referencia a los GAE de cada organismo. Para cada especie se calcularon 1.000 valores de S, a partir de genes elegidos en forma aleatoria del genoma. Se registró el rango de los valores de S incluyendo el 95% de las muestras. El valor de 0,4 se estableció como umbral para considerar significativa la selección sobre la traducción. La inercia filogenética asociada con el coeficiente de selección en *Comamonadaceae* se estimó siguiendo el método desarrollado por Vieira-Silva y Rocha (2008). Para cada par de genomas, se estudió la asociación entre la distancia cofenética y el valor absoluto de la diferencia del coeficiente de selección. La distancia cofenética fue estimada en base al largo de las ramas de árbol filogenético utilizando la función cophenetic en el paquete APE en R (Paradis et al. 2004). Se realizó la reconstrucción ancestral del coeficiente de selección utilizando al función fastAnc en el paquete phytools (Revell, 2012).

El estudio exhaustivo de la composición de las secuencias codificantes se hizo utilizando el programa CodonW 1.4.2. Con el mismo programa se estimó para cada gen las frecuencias de bases en la primera, segunda y tercera posición del codón y la frecuencia de guanina más citosina (GC). Para la estimación de las bases GC en la tercera posición del codón se utilizaron a las secuencias de los codones pertenecientes a los cuartetos naturales y a los cuartetos de los sextetos. Se predijo la presencia de los genes codificantes para los ARNt y sus respectivos anticodones en cada genoma,

utilizando el programa tRNAscan-SE-1.3.1, aplicando un modelo "*mixed*" general con parámetros establecidos por defecto (Lowe y Eddy 1997) (http://lowelab.URCSc.edu/tRNAscan-SE/).

3.6 Estimación de la distancia molecular

Los secuencias ortólogas entre los organismos de la familia *Comamonadaceae*, tomados de a pares, fueron identificadas utilizando el criterio de "*best reciprocal hit*" como se describió previamente en la sección "Reconstrucción filogenética". Las secuencias obtenidas fueron traducidas y alineadas utilizando el programa ClustalW (Thompson et al. 1994); las secuencias de ADN se alinearon a partir del alineamiento aminoacídico utilizando el programa Tranaling implementado en el paquete EMBOSS (Rice et al. 2000). Las distancias moleculares sinónimas (dS) y no sinónimas (dN) fueron estimadas en base a un modelo de máxima verosimilitud, utilizando el programa codeml, que se incluye en el paquete PAML4.0 (Yang 2007). Sólo se consideraron secuencias con una divergencia sinónima menor a 1,5 ($dS \leq 1,5$).

3.7 Sesgo en el uso de codones en las regiones conservadas en las proteínas

Las regiones conservadas (RC) y no conservadas (NRC) fueron identificadas y separadas utilizando el programa Gblock con los siguientes parámetros –b1=32, –b2=32, –b3=1 and –b4=2 (Talavera y Castresana 2007). Bloques con una longitud de al menos dos aminoácidos sin cambios en todas las especies y con un máximo de una posición contigua no conservada, fueron agrupados y utilizados para este estudio. Lo robusto de este método para definir las RC fue testeado utilizando estimaciones de p-distancias pareadas, utilizando el programa distmat implementado en el paquete EMBOSS (Rice et al. 2000). El valor de p-distancia es proporcional a los sitios de aminoácidos en la que las dos secuencias comparadas son diferentes. El sesgo en el uso de codones en las RC fue estimado como la diferencias entre el RSCU de cada codón entre las RC respecto a las NRC. Se utilizó el estadístico χ^2 de contingencia para determinar que codón se utiliza en forma significativamente más frecuente en las RC.

3.8 Predicción de los genes de las proteínas de membrana

Mediante el programa TMHMM basado en un modelo de cadenas ocultas de Markov (disponible en la página web http://www.cbs.dtu.dk/services/TMHMM/), se identificaron los genes que codifican

proteínas de membrana (Krogh et al. 2001; Sonnhammer et al. 1998). A partir del resultado del programa TMHMM se identificaron las regiones TM y NOTM en dichas secuencias, las que fueron obtenidas, concatenadas y clasificadas por un script realizado en el laboratorio. Se utilizó el estadístico χ^2 de contingencia para determinar los codones más utilizados por las regiones TM con respecto a las regiones NOTM.

3.9 Estructura secundaria del ARN mensajero

Se calculó la energía mínima libre (MEL) utilizando el programa RNAfold, incluido en el paquete ViennaRNA versión 2.1.5 (Lorenz et al. 2011; Zuker y Stiegler 1981). Los valores de MEL son expresados como negativos y en unidades Kcal/mol. Para poder hacer comparables a los valores de MEL de las distintas secuencias de ARNm, se normalizó por la longitud de las secuencias (MELN). MELN se calcula dividiendo el valor de MEL por la longitud de la secuencia del ARNm y multiplicando el resultado por 100 para relacionar el valor de MEL con un segmento de 100 nucleótidos: MELN=100*MEL/largo (Zhang et al. 2006).

Para poder tener variables estadísticamente independientes, o sea sin la influencia de la inercia filogenética, se estimaron los valores para la variables de interés utilizando el método phylogenetic independent contrast (PIC) (Felsenstein 1985) utilizando el paquete ape (Paradis et al. 2004) del programa R (Development Core Team 2011), tomando en cuenta la filogenia inferida utilizando a las proteínas ribosomales.

RESULTADOS Y DISCUSIÓN.

4.1 Ensamblado del genoma y análisis composicional de Delftia sp. JD2

La secuenciación del genoma de *Delftia sp* JD2 presentó un total de 15:287.096 "paired-end reads" (2 X 100 ciclos) con una cobertura de 36,8. Después de realizado el ensamblado *de novo*, se obtuvo un total de 219 contigs con un longitud promedio de 30.894 pb (274.616 pb el más largo y 77 pb el más corto). Este resultado sugiere que *Delftia sp.* JD2 presenta un cromosoma circular con un total de 6:765.786 pb de largo, y su contenido en GC es de 67%. La anotación del borrador del genoma se realizó utilizando RAST mediante el subsistema de SEED. El genoma predicho presenta 6.051 secuencias codificantes para proteínas y dos copias del ARNr16S. Entre los genes predichos que codifican proteínas, 25,7% son consideradas como hipotéticas y 2.647 genes fueron clasificados en 500 subsistemas por RAST. Para entender los procesos que actúan en el genoma de *Delftia sp.* JD2 haciendo énfasis en aspecto evolutivos, se realizó una comparación entre éste genoma y los de *Delftia acidovorans* SPH-1 y *Delftia* Cs1-4.

Utilizando las secuencias de los genomas de *Delftia* Cs1-4 y *D. acidovorans* SPH-1 con respecto al genoma de *Delftia sp.* JD2 se estimaron los valores de ANI, los que fueron de 98,65% o 98,07% entre *Delftia sp.* JD2 y *D. acidovorans* SPH-1 o *D.* Cs1-4, respectivamente. Por lo general, una gama umbral de 95-96% se considera una medida robusta de similitud genómica entre cepas (Kim et al. 2014). Este resultado sugiere que estas cepas probablemente pertenezcan a la misma especie que *D. acidovorans* SPH-1.

Mediante un script realizado en el laboratorio se identificaron los genes ortólogos entre las tres *Delftias*, siguiendo el criterio de "Best Reciprocal Hit" entre las proteínas, utilizando la herramienta BLASTp (Altschul et al. 1990). A partir de los genes ortólogos se analizó el pan-genoma entre las bacterias *Delftia sp.* JD2, *D. acidovorans*. SPH-1 y *Delftia* Cs1-4 (Figura 1). El diagrama de Venn de los tres organismos presenta un genoma "core" compuesto por 4.799 genes ortólogos, correspondiendo aproximadamente al 79% de todas las secuencias codificantes de proteínas del genoma del *Delftia sp.* JD2 (Figura 1). El alto número de genes que comparten los tres genomas refleja una alta conservación a nivel de secuencia y función dentro del género. La cepa *Delftia sp.* JD2 presentó un total de 750 genes que no pudieron cumplir con el criterio de "Best Reciprocal Hit", aproximadamente el 12,4% de las secuencias codificantes de proteínas. Estos genes supuestamente sólo están presentes en ella, no presentando ortólogos en las otras dos cepas.

Naturalmente, con la disponibilidad de nuevas secuencias de Delftia, estos valores podrían cambiar.

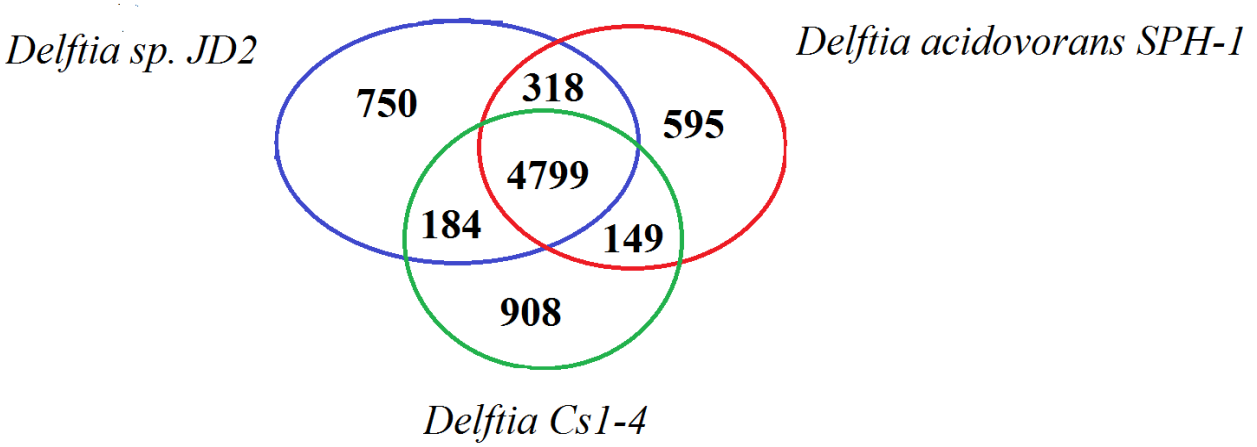

Figura 1. Diagrama de Venn de los genes ortólogos en las tres *Delftia* muestra el pan genoma de las cepas. Las regiones solapadas representan a los genes ortólogos en común entre los genomas.

4.2 Análisis de posibles eventos de transferencia horizontal de genes en Delftia sp. JD2

Se identificaron 301 genes que se encuentran únicamente en *Delftia sp.* JD2 con respecto a las otras dos cepas completamente secuenciadas. Los genes únicos presentaron una media de MELP de 0,84 ± 0,22, mientras que los genes que codifican a las proteínas ribosomales (consideradas de alta expresión) presentan valores más altos con una media de 1,7 ± 0,34 (Figura 2). Por otra parte, los genes ortólogos son el grupo que presentan en promedio el valor de MELP más bajo: 0,72 ± 0,18 (Figura 2). Los tres grupos de genes presentan diferencias significativa entre sí (W, $p<0,05$). Los genes únicos presentan valores superiores con respecto a los valores que presentan los genes ortólogos, sin embargo no llegan a alcanzar los valores que presentan los genes que codifican a las proteínas ribosomales. A partir de este resultado se puede inferir que los genes únicos no son genes de alta expresión, sin embargo presentan un valor superior a la media del índice de expresión (MELP) con respecto a los genes ortólogos (Figura 2).

Al comparar el contenido de GC entre los tres grupos de secuencias, se puede apreciar que los genes que codifican a las proteínas ribosomales con respecto a los genes únicos no presentan diferencias significativas (W, $p> 0,05$). Los genes ribosomales y los genes únicos presentan un valor medio de GC de 0,58 ± 0,02 y 0,60 ± 0,02 respectivamente (Figura 2). Por otra parte, los genes ortólogos presentan un media en el contenido de GC de 0,67 ± 0,03. De esta manera, los genes ortólogos

presentan un enriquecimiento significativo (W, $p < 0,05$) en el contenido de GC, con respecto a los genes únicos y a los genes ribosomales. Los genes únicos podrían ser consecuencias de transferencia horizontal de genes (THG) ya que no solamente no están presentes en cepas estrechamente emparentadas, si no que a su vez presentan un patrón diferente al de los genes ortólogos, con respeto al contenido de GC.

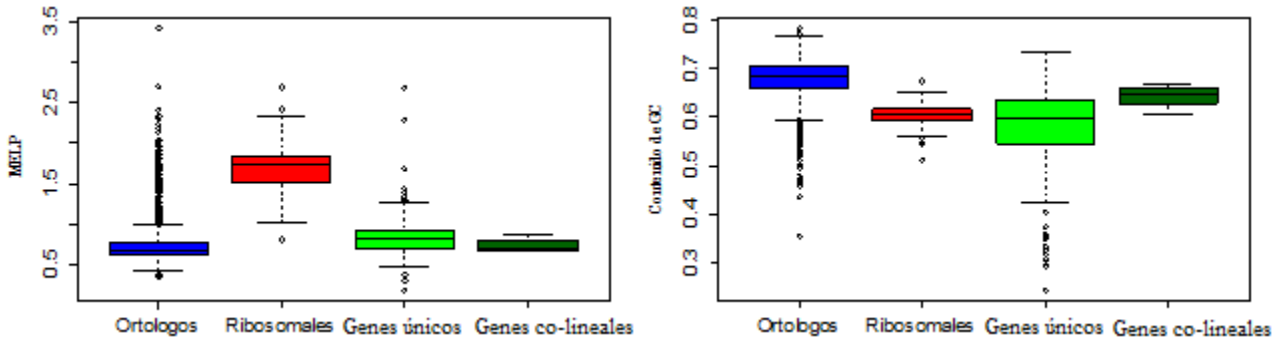

Figura 2. Se muestra la distribución de los valores de MELP y el contenido de GC de los genes ortólogos entre las tres bacterias del género Delftia, de *Delftia sp.* JD2, los genes de las proteínas ribosomales y los genes que se encuentran solamente en la bacterias *Delftia sp.* JD2 con respecto a las otras dos cepas. Se muestran a los ocho genes que serían probablemente el resultados de posibles transferencia horizontal, y que presentan colinealidad entre *Delftia sp.* JD2 con organismos filogenéticamente lejanos.

Se ha demostrado que *Delftia sp.* JD2 presenta características biológicas importantes, por ejemplo se la puede considerar como una bacteria promotora de crecimiento de plantas, presentando a su vez genes de resistencia a metales pesados como por ejemplo al cromo (Morel et al. 2011). De allí resulta plausible postular que la incorporación de genes por transferencia horizontal puede, en estos genomas, estar vinculada con la riqueza metabólica que demuestran. Como se ha mencionado anteriormente, uno de los factores que impulsan la expansión de los genes en los procariotas es la transferencia horizontal (Treangen y Rocha 2011).

Entre los 301 genes únicos que presentan *Delftia sp.* JD2 existen varios genes con funciones de resistencia a metales pesados. Uno de ellos tiene codifica para la resistencia a cobalto, zinc y cadmio (dss.5102, Cobalt-zinc-cadmium resistance protein CzcD). Se construyó una filogenia tomando a ese gen y las secuencias más similares, en una búsqueda a través de un blastp tomando a todos los genomas y plásmidos disponibles en el Genbank (Marzo 2016). A partir de la filogenia, se puede apreciar que el gen dss.5102 se asocia con los organismos *Pseudomonas aeruginosa* LESB58, *Parvibaculum lavamentivorans* DS-1 y *Cupriavidus metallidurans* CH34 con un gran apoyo en el nodo (Figura 3). Las bacterias que se asocian con *Delftia sp.* JD2, provienen de grupos lejanos filogenéticamente y todas están incluidas en el *phylum* Proteobacterias. *P. aeruginosa* LESB58 y *P. lavamentivorans* DS-1 están clasificadas dentro de la clase Gammaproteobacteria y Alphaproteobacteria, respectivamente. En cuanto a la bacteria *C. metallidurans* CH34 es la que se

encuentran más cercana con respecto a *Delftia sp.* JD2 ya que ambas pertenecen a la clase Betaproteobacteria, aunque en diferentes familias.

Se encontró una colinealidad de ocho genes que incluyen a dss.5102 en *Delftia sp.* JD2, *P. aeruginosa* LESB58, *P. lavamentivorans* DS-1 y *C. metallidurans* CH34 (Tabla 1). A pesar de ser organismos filogenéticamente lejanos, esta región de ocho secuencias está constituida por genes con una función similar entre sí, pudiendo estar involucrados en una misma ruta metabólica (Zaslaver et al. 2006), lo que sugiere que se podría tratar de un operón (Wolf et al. 2001). Tomando en cuenta las predicciones de los operones realizadas por Taboada et al. (2012), se pudo identificar que dos de estos genesson parte de un operón mientras que los otros genes no están organizados de esta forma en los organismos *P. lavamentivorans* DS-1 y *C. metallidurans* CH3 (Tabla 1). En cambio *P. aeruginosa* LESB58 presenta cuatro genes organizados en dos operones y las otras secuencias no están incluidos en este tipo de organización. Cuando se analizó a bacterias cercanas filogenéticamente como es *Delftia* Cs 1-4 no se encontró ninguno de estos genes. En cambio, cuando se estudió a *D. acidovorans* SPH-1 se encontró que ésta presenta siete de los ocho que presentaron colinealidad, con una función similar, aunque hay que destacar que en esta última especie los genes no están ubicados en forma contigua en el genoma. Efectivamente, tres de estos genes se sitúan en una misma región (y dos de ellos constituyen un operón), hay otra secuencia aislada mientras que los cuatro genes restantes se encuentran en otra región del genoma sin estar organizados como un operón (Tabla 1). Dado que estos genes están involucrados en la misma función (resistencia a metales pesados), y asumiendo que es muy poco probable que el mismo operón se genere más de una vez en forma independiente (Lawrence 1999; Lawrence 1997), y más aún, es un evento improbable que ocho secuencias se organicen en forma colineal por azar en organismos filogenéticamente lejanos, es que postulamos que estas secuencias aparecen en *Delftia sp.* JD2 como resultado de un evento de transferencia horizontal.

Tabla1.Colinealidad entre los genes ortólogos de las bacterias *Delftia sp* JD2 en los organismos *Delftia acidovorans* SPH-1, *Cupriavidus metallidurans* CH34, *Parvibaculum lavamentivorans* DS-1 y *Pseudomonas aeruginosa* LESB58. Se muestra la posición de los genes en el genoma, la descripción de la función de los genes y en la columnas de las posición de los genes en el operón los que están coloreados con gris pertenecen a un operón y por último se muestran en qué posición del operón se encuentra.

Posición de los genes en el genoma	Función del gen	Posición de los genes en el genoma	Función del gen	Posición del gen en el operón
Delftia sp. JD2		*Delftia acidovorans* SPH-1		
5099	Lipoprotein signal peptidase (EC 3.4.23.36)	479	lipoprotein signal peptidase	1
5100	Lead, cadmium, zinc and mercury transporting ATPase	478	heavy metal translocating P-type ATPase	2
5101	Transcriptional regulator, MerR family Cobalt-zinc-cadmium resistance	477	MerR family transcriptional regulator	1
5103	DNA topoisomerase III	4134	DNA topoisomerase III	1
5104	Single-stranded DNA-binding protein in PFGI-1-like cluster	4133	single-stranded DNA-binding protein	1
5105	Integrase regulator R	4132	putative integrase regulator R protein	1
5106	hypothetical protein	4131	hypothetical protein	1
Delftia sp. JD2		*Cupriavidus metallidurans* CH34		
5099	Lipoprotein signal peptidase (EC 3.4.23.36)	2365	prolipoprotein signal peptidase (signal peptidase II)	2
5100	Lead, cadmium, zinc and mercury transporting ATPase	2364	lead/cadmium-transporting ATPase	1
5101	Transcriptional regulator, MerR family Cobalt-zinc-cadmium resistance	2363	MerR family transcriptional regulator	1
5102	Cobalt-zinc-cadmium resistance protein CzcD	2358	cation efflux system protein	1
5103	DNA topoisomerase III	2357	DNA topoisomerase III	1
5104	Single-stranded DNA-binding protein in PFGI-1-like cluster	2356	single-stranded DNA-binding protein	1
5105	Integrase regulator R	2355	Integrase regulator R	1
5106	hypothetical protein	2354	hypothetical protein	1
Delftia sp. JD2		*Parvibaculum lavamentivorans* DS-1		
5099	Lipoprotein signal peptidase (EC 3.4.23.36)	3371	lipoprotein signal peptidase	2
5100	Lead, cadmium, zinc and mercury transporting ATPase	3370	lipoprotein signal peptidase	1
5101	Transcriptional regulator, MerR family Cobalt-zinc-cadmium resistance	3369	MerR family transcriptional regulator	1
5102	Cobalt-zinc-cadmium resistance protein CzcD	3367	cation efflux protein	1
5103	DNA topoisomerase III	3366	DNA topoisomerase III	1
5104	Single-stranded DNA-binding protein in PFGI-1-like cluster	3365	single-stranded DNA-binding protein	1
5105	Integrase regulator R	3364	hypothetical protein	1
5106	hypothetical protein	3363	hypothetical protein	1
Delftia sp. JD2		*Pseudomonas aeruginosa* LESB58		
5099	Lipoprotein signal peptidase (EC 3.4.23.36)	2677	putative lipoprotein signal peptidase LspA	2
5100	Lead, cadmium, zinc and mercury transporting ATPase	2678	Heavy metal translocating P-type ATPase	1
5101	Transcriptional regulator, MerR family Cobalt-zinc-cadmium resistance	2679	transcriptional regulator, MerR family	1
5102	Cobalt-zinc-cadmium resistance protein CzcD	2680	Co/Zn/Cd efflux system protein	1
5103	DNA topoisomerase III	2681	DNA topoisomerase III	1
5104	Single-stranded DNA-binding protein in PFGI-1-like cluster	2682	single-stranded DNA-binding protein	1
5105	Integrase regulator R	2683	Putative integrase regulator R protein	2
5106	hypothetical protein	2684	hypothetical protein	1

Se analizó el contenido de GC de los ocho genes que habrían sido obtenidos por transferencia horizontal. Estos mostraron una media de 64% (± 2%), valor que es más alto al contenido de GC de los genes que codifican a las proteínas ribosomales y los genes que se encuentran únicamente en *Delftia sp.* JD2 (301 secuencias). Sin embargo, presentan un contenido inferior de GC con respecto a los genes ortólogos. El grupo de ocho genes presenta diferencias significativas en su contenido de GC en todas las comparaciones (W, $p<0,05$) (Figura 2). A partir de este resultado se podría inferir que estas secuencias que se presume que hayan sido obtenidos por transferencia horizontal de genes, sean originarios de un organismo que presenta, en promedio, un contenido de GC inferior al que presenta *Delftia sp.* JD2.

Por otra parte, los ocho genes tienen una media en el valor de MELP de 0,73 ± 0,09, lo cual no representa una diferencia significativa (W, $p>0,05$) con respecto a los genes ortólogos y los genes únicos. Pero sí presenta una diferencia significativa (W, $p<0,05$) cuando se los testeó contra la distribución de los valores de MELP de las proteínas ribosomales. Por lo tanto estos ocho genes se puede considerar que no son genes de alta expresión.

Como indicábamos más arriba, proponemos que estos ocho genes que están presenten en *Delftia sp.* JD2 sean probablemente la consecuencia de un evento de transferencia horizontal. La idea más parsimoniosa para poder explicar esto, sería que el ancestro común de *Delftia sp.* JD2 y *D. acidovorans* SPH-1 los hayan adquirido y mientras que en *D. acidovorans* SPH-1, por la dinámica del genoma, se separaron en *Delftia sp.* JD2 se mantuvieron contiguos en una sola región del genoma. Un solo organismo de la familia *Comamonadaceae* que es *Acidovorax ebreus* TPSY se asocia a la filogenia realizada por el gen dss.5102, ningún otro organismo de esta familia está presente en la filogenia (Figura 3). De este resultado se puede inferir que lo más probable es que se hayan ganado estos genes por los organismos que lo poseen *A. ebreus* TPSY, *Delftia sp.* JD2 y *D. acidovorans* SPH-1, mientras que lo perdieron los otros organismos de la familia.

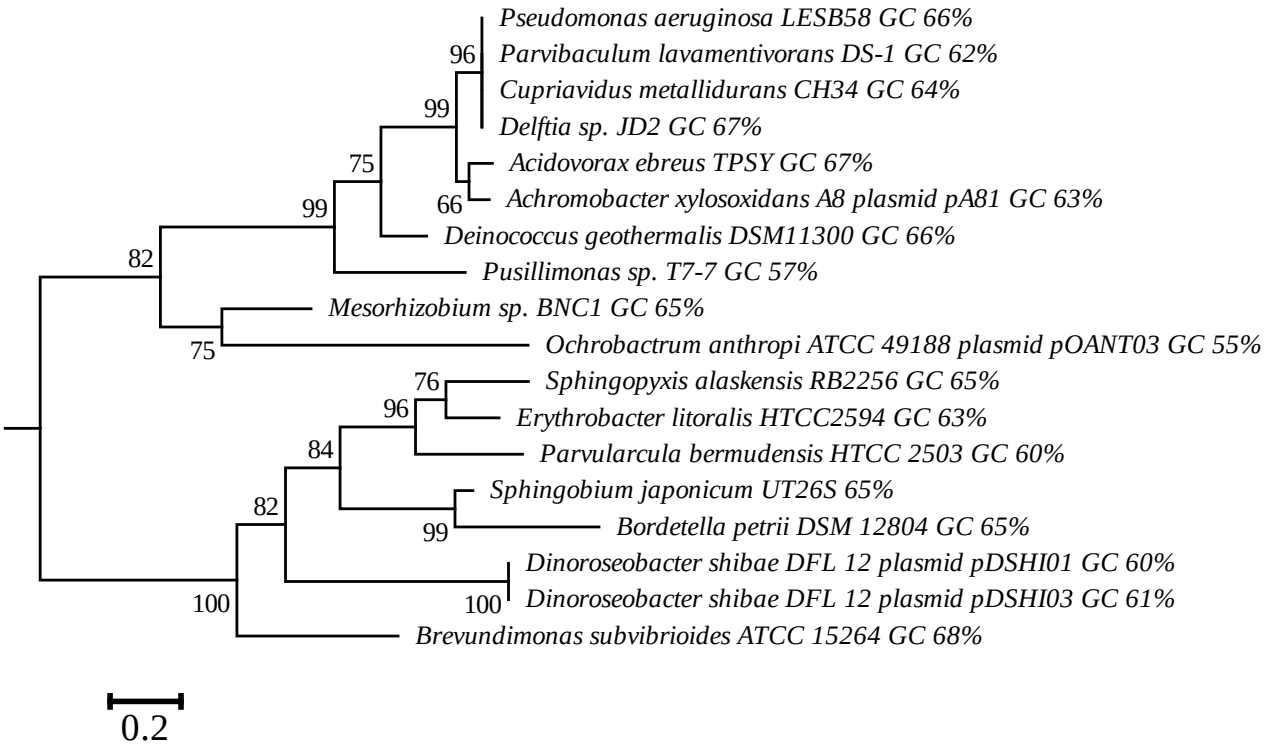

Figura 3. Reconstrucción filogenética de los organismos que "pegaron" mejor mediante el blastp respecto al gen de Delftia sp JD2 dss.5102, mediante el programa PHYML.

4.3 Análisis filogenético y contenido de GC

La reconstrucción filogenética fue realizada a partir de las proteínas ribosomales. El árbol de la filogenia concuerda con reconstrucciones previas para este grupo (Chen et al. 2012; Leadbetter y Greenberg 2000; Wen et al. 1999) (Figura 4). Los organismos pertenecientes al género *Acidovorax* aparecen en la filogenia como un grupo polifilético, ya que algunas cepas se agrupan con las bacterias *Verminephrobacter*, *Delftia*, *Comamonas* y *Alicyliphilus*. Las bacterias que pertenecen al género *Acidovorax* están distribuidas en tres grupos en el árbol. Dos de los tres grupos son *Acidovorax* viven en forma libre: *A. radicis*, *Acidovorax* CF316, *Acidovorax* NO-1, *Acidovorax* KKS102 y *A. delafieldii*; el segundo grupo está constituido por *Acidovorax* JS42 y *A. ebreus* YPSY, mientras que el tercero está constituido por las bacterias patógenas de plantas *A. avenae* ATCC 19860 y *A. citrulli* AAC00-1. El grupo que está constituido por las *Acidovorax* de vida libre tiene como vecinos a las *Comamonas*, *Delftia*, *Alicyliphilus* y *Verminephrobacter* (Willems 2014).

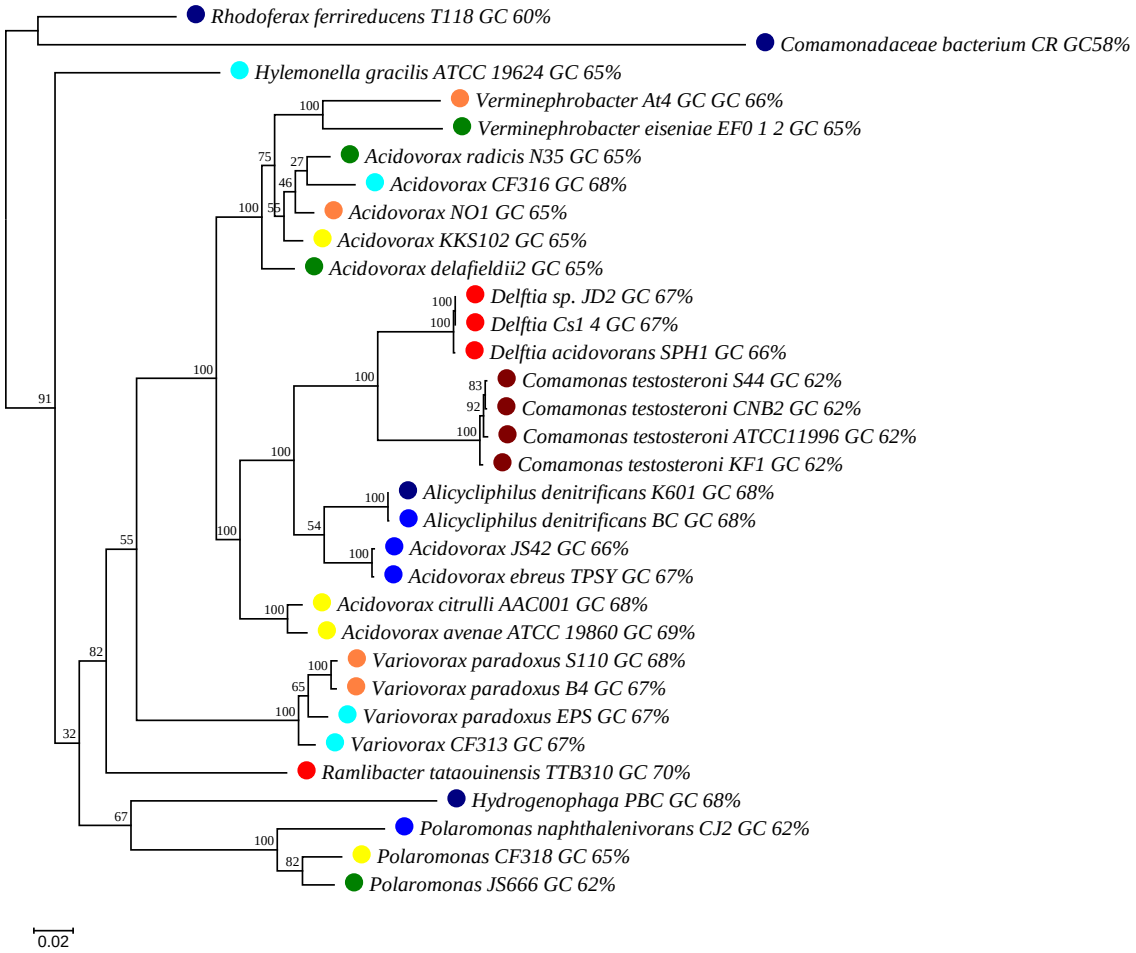

Figura 4. Reconstrucción filogenética tomando a las proteínas ribosomales. Los organismos que tiene el redondel verde, presentan diferencia significativa en el valor de S.

El rango en el contenido de GC genómico de la familia *Comamonadaceae* va desde un 58% a un 70% en el organismos *Comamonadaceae bacterium* CR y *Ramlibacter tataouinensis* TTB310 respectivamente (media=65,4; SD=2,8%) (Tabla 2). Se ha estudiado la distribución en el contenido de GC para cada posición del codón en los genomas completos y en los genes ortólogos (Figura A1). La distribución del contenido de GC para cada posición del codón es muy similar entre los genomas analizados y entre los genes ortólogos (Figura A1). Como regla general el contenido de GC3 es más alto en comparación con el contenido de GC, llegando a presentar una media cercana a 90%. Este comportamiento podría estar reflejando el sesgo mutacional de todos los organismos analizados. Este sesgo es más pronunciado en cepas de *R. tataouinensis* TTB310 y *Alicycliphilus*. Sin embargo, organismos como *C. bacterium* CR o *C. testosterone* presentan un contenido de GC más bajo en esa posición. En todos los organismos, la distribución del contenido de GC para cada posición de los codones entre los genes ortólogos y el genoma completo no presentan diferencias significativas (W, p> 0,05), la diferencia de estos dos grupos de genes es despreciable (diferencias de medias = 1%, SD=0.7%, Tabla A1). Como regla general, de este resultado podemos inferir que

los genes ortólogos pueden considerarse representativos de la tendencia composicional de bases del genoma (Tabla A1).

Se pudo relacionar el contenido de GC de los organismos en estudio con la distancia filogenética (Figura 5), observándose que los organismos más cercanos entre sí presentan un contenido de GC más similar.

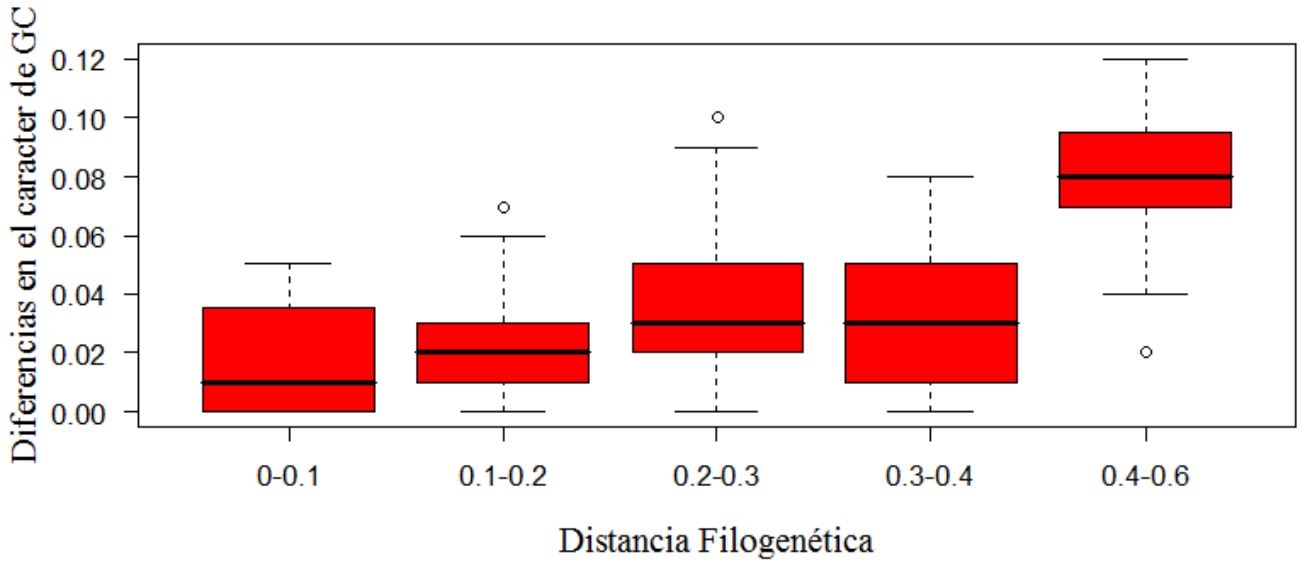

Figura 5. Se muestra la correlación entre la distancia cofenética y la diferencia en el contenido de GC en valor absoluto entre los organismos. Se puede observar que existe una tendencia a que los organismos cercanos filogenéticamente presentan a su vez valores similares en el contenido de GC, mientras que en los organismos que se encuentran distantes presentan el comportamiento opuesto.

4.4 Análisis del sesgo en el uso de codones sinónimos.

4.4.1 Análisis multivariado

Con el fin de poder estudiar la variabilidad observada en el uso de los codones sinónimos en cada especie, se realizó un análisis multivariado (ver Materiales y Métodos). La varianza explicada por el primer eje es relativamente alta en todos los organismos variando entre un 21,6 hasta 32,3%, mientras que el segundo eje presenta una variabilidad entre un 3,9 a 10,6% (Tabla 2).

En los organismos *Comamonadaceae bacterium* CR, *Rhodoferax ferrireducens* T118, *Verminephrobacter eiseniae* EF01 2, *Hylemonella gracilis* ATCC 19624, *Verminephrobacter* At4, los genes de alta expresión (GAE) se agrupan hacia un extremo del primer eje (Tabla 2, Figura A2). En el resto de los organismos los GAE se agrupan hacia un extremo del segundo eje (Tabla2, Figura A2). En los organismos que se asocia el segundo eje del análisis WCA con el MELP, presentan una baja variabilidad, esto sugiere que el sesgo en el uso de codones sinónimos varía poco a través de

estos genomas. Por otro lado, los Genes de Baja Expresión (GBE) y genes que codifican a proteínas hipotéticas están distribuidos en forma aleatoria en todo el eje. Se ha encontrado una correlación significativa entre los dos primeros ejes del análisis WCA, con el valor del índice de expresión (MELP) para cada gen y su posición a lo largo del eje (Tabla 2). El valor de correlación entre el primer o segundo eje generado por el WCA con el indicie de expresión varía entre organismos ($0,88 \geq r \geq 0,37$; $p<0,05$) (Tabla 2). Se ha encontrado que el porcentaje total de variabilidad explicado por los ejes asociados a nivel de expresión muestra que es alto en algunas especies y en otras no, lo que indica que existe variabilidad entre las especies (Tabla 2).

Este resultado sugiere que la expresión es uno de los principales factores que puede explicar la variación en el CU intra-genómica en todas las especies. Se puede sugerir que el sesgo en el uso de codones sinónimos está asociado con el nivel de expresión, variando en los organismos en estudio.

4.4.2 Análisis del coeficiente de selección S (Sharp).

Se ha comparado el efecto de la selección actuando al nivel de la velocidad de traducción utilizando la intensidad del sesgo en el uso de codones sinónimos (S) desarrollado por Sharp et al. (2005). Los valores de S fueron estimados comparando a los GEA con respecto a los GBE, asumiendo que la frecuencia del uso de los codones sinónimos en los GBE son generados como el resultado del sesgo mutacional en ausencia de selección (Chen et al. 2004). En cambio en los GAE se asume que la selección presentan una mayor acción con respecto a los GBE (Goetz y Fuglsang 2005). Como regla general, los organismos de vida libre con tiempo corto de duplicación, como por ejemplo *Escherichia coli* K12, *Clostridium perfringens* CT y *Bacillus subtilis*, se caracterizan por presentar altos valores del coeficiente de selección, 1,49; 2,65 y 1,36, respetivamente (Sharp et al. 2005). Sin embargo, organismos parásitos intracelulares o endosimbiontes como *Rickettsia prowazekii*, *Buchnera aphidicola*, *Wigglesworthia glossinidia*, entre otros tienden a presentar valores bajos de S (< 0.2) (Sharp et al. 2005).

Los valores de S presentaron una amplia variación entre las especies, desde -0,20 en las bacteria *Comamonadaceae bacterium* CR a 1,18 en *Comamonas testosteroni* CNB 2 (Tabla 2). Para 17 especies, el valor de S es más alto al límite superior del 95% de los genes elegidos en forma aleatoria y mayor a 0,4 (Tabla 2). A partir de este resultado se puede inferir que la selección podría estar afectando al uso de codones en esos genomas. Por otro lado, seis especies presentan un valor de S inferior al límite superior del 95% para los genes elegido en forma aleatoria y menores a 0,4. Por último, para nueve especies el valor de S no supera a 0,4 pero sí presentan valores más altos al

límite superior del 95% de los genes elegidos en forma aleatoria (Tabla 2). De esta manera, no hay evidencia clara de que la selección esté afectando al uso de codones a nivel de velocidad en estos genomas. En el trabajo de Sharp et al. (2005), se analizaron tres genomas de Betaproteobacteria, presentando valores bajos de S, aunque ninguno de estos tres pertenecen a la familia Comamonadaceae. La comparación del valor del coeficiente de selección entre diferentes familias de otras clases, muestra que los valores de S de los organismos en estudio nos es alta como se esperaría considerando la variabilidad eco fisiológica (Iriarte et al. 2013; Iriarte et al. 2014). Los géneros *Delftia* y *Comamonas* muestran los valores más altos de S, que van desde 0,60 hasta 1,18 (Tabla 2). Este resultado sugiere que existe una fuerte o moderada selección impulsando el sesgo en el uso de codones en estos géneros. Por otra parte, organismos pertenecientes a los géneros *Polaromonas*, *Verminephrobacter* y *Variovorax* presentan un bajo coeficiente de S, pero significativo. El género *Acidovorax* aparece como un grupo polifilético, y el patrón de los valores de S no es conservado dentro del grupo. Las especies *A. ebreus* TPYS, *Acidovorax* JS42, *Acidovorax* CF316 y *A. radicis* N35 presentan un valor de S de 0,15 a 0,39, estas no cumplen con nuestras condiciones para ser consideradas que estén bajo selección. Sin embargo *A. delafieldii* 2, *A. citrulli* AAC00 1, *A. avenae* ATCC 19860, *Acidovorax* KKS102 y *Acidovorax* NO 1 tienen un S relativamente alto y significativo con un rango de 0,40 a 0,51 (Tabla 2).

Por otro lado, las especies *Comamonadaceae bacterium* CR, *Hydrogenophaga* PBC, *Rhodoferax ferrireducens* T118, *Alicycliphilus denitrificans* BC, *A. denitrificans* K601, presentan valores de S de -0,20 a 0,14. Estos valores son bajos y no significativos, lo que sugiere que existe una selección suave o no existe selección en cuanto al sesgo en el uso de codones sinónimos en estos organismos.

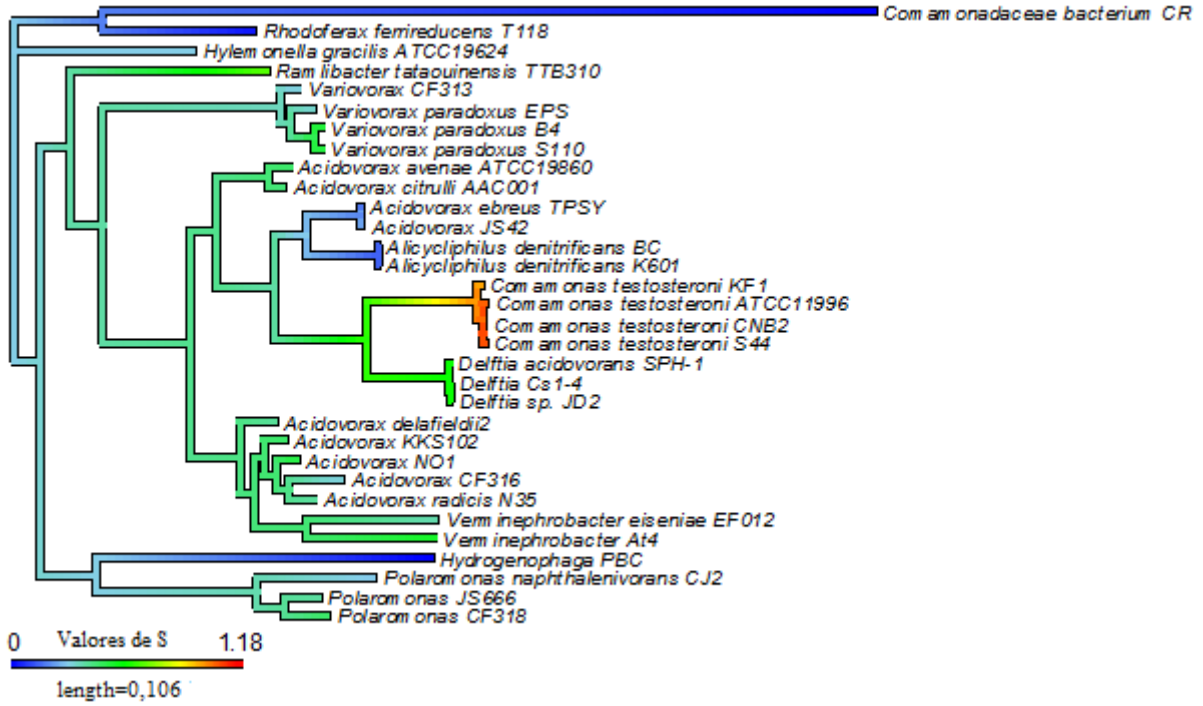

Figura 6. Reconstrucción ancestral del coeficiente de selección (S).

La reconstrucción ancestral del coeficiente de selección demostró que la selección en el uso de codones sinónimos es relativamente reciente, ya que presenta valores moderados (Figura 6). Es importante destacar que valores de S altos o moderados son identificados en el grupo de los géneros *Delftia* y *Comamonas*, así como también en el organismo *Ramlibacter tatouinensis* TTB310. Al comparar los codones óptimos y el pool de ARNt entre los organismos de los géneros *Delftia* y *Comamonas* con respecto a *R. tatouinensis* TTB310, encontramos que son, en general, diferentes. De esta manera la selección entre estos organismos puede que sea un proceso paralelo. Este comportamiento, puede ser explicado parcialmente por la hipótesis de "relajación de la selección" (Sharp et al. 2010). Esta hipótesis propone que la selección debe ser débil en un período prolongado de tiempo. De esta forma al relajarse la selección a nivel de la traducción, abre la posibilidad de generar nuevos patrones en la elección de los codones y a su vez en el pool de ARNt. Luego cuando la selección haya recuperado su fuerza, se generará un nuevo estado adaptativo entre los nuevos grupos de codones preferidos y el pool de ARNt.

Asimismo, el sesgo en la selección parece estar fuertemente influenciado por la inercia filogenética, ya que los organismos que se encuentran cercanos en la filogenia presentan valores de S similares, mientras que en los organismos distantes presentan valores de S diferentes (Figura 7). La selección en el sesgo en el uso de codones es muy variada en esta familia, presentando valores significativos pero moderados en la mayoría de los organismos. Sin embargo, las tendencias asociadas a la expresión en el sesgo del uso de codones sinónimos identificado por el análisis de WCA pueden ser el resultado de una combinación de múltiples factores, incluyendo la selección y el sesgo mutacional, siendo el último mencionado como dominante en el caso de los organismos con bajos valores de S.

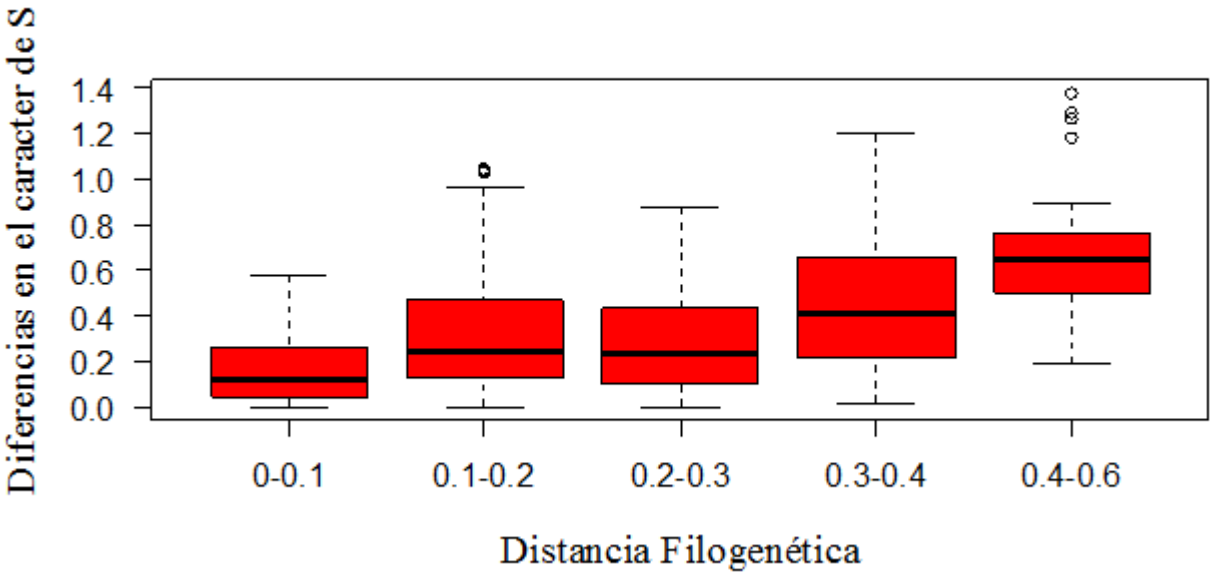

Figura 7. Se muestra la correlación entre la distancia cofenética y la diferencia en valor absoluto entre los organismos del índice de Sharp (S). Se puede observar que existe una tendencia a que los organismos cercanos filogenéticamente presentan a su vez valores similares de S, mientras que en los organismos que se encuentran distantes presentan el comportamiento opuesto.

4.4.3 Selección en la exactitud ("fidelidad") de la traducción

La incorporación errónea de un aminoácido puede afectar la adaptabilidad (el "fitness") y se ha demostrado que cambios sinónimos pueden alterar la frecuencia de incorporaciones aminoacídicas erróneas a la proteína (Precup y Parker 1987). También se ha establecido que los codones sinónimos presentan diferentes tasas de error. Como resultado de lo anterior, se espera que la selección opere a nivel de la fidelidad (sobre todo en regiones conservadas), determinando un uso particular de codones sinónimos. Esto fue demostrado en primer lugar por Akashi (1994) y confirmado posteriormente en varios trabajos (por ejemplo, Stoletzki y Eyre-Walker 2007).

A partir de los 601 genes ortólogos el 40% de las secuencias codificantes se agruparon como RC. La p-distancia fue estimada para cada gen tomando en cuenta a las secuencias conservadas y no conservadas. En promedio las RNC fueron 14 veces más distantes en comparación con las regiones conservadas, de esta manera se valida el criterio para poder elegir las regiones conservadas (Figura A3). Las RC presentaron entre 15 a 20 codones utilizados significativamente con más frecuencia en los 17 organismos que presentan selección (S>0,4; significativos) y tienden a ser los más utilizados en los genes de alta expresión. Es de destacar que entre ellos la coincidencia varía de un 82% hasta el 89% con el ARNt isoaceptor presente en el genoma (Figura 8).

Se identificó una correlación positiva entre el sesgo en el uso de codones entre las RC y los GEA en casi todos los organismos, con la excepción de *C. bacterium* CR, y fueron significativas en 21 organismos ($0.26 < r < 0.73$, $p < 0,05$) (Tabla 2). Es importante destacar que los organismos que presentan valores altos de S, como por ejemplo *R. tataouinensis* TTB310, *Delftia*, *Comamonas* y *Verminephrobacter*, presentan también altos valores de correlaciones, sugiriendo que cuando la selección opera en el sesgo en el uso de codones en los genes de alta expresión, los mismos codones tienden a ser los preferidos en las regiones conservadas.

Por lo tanto, se concluye que existe selección en el uso de codones en las RC en esta familia. La selección puede favorecer a los codones que se traducen con mayor precisión. A pesar de ser variables, esta tendencia observada puede ser considerada como un rasgo conservado en la familia.

4.4.4 Análisis e identificación de los codones óptimos y rechazados por los genes que codifican a las proteínas ribosomales, asociaciones con el ARNt.

Los codones óptimos fueron identificados en los organismos que presentan selección o sea un valor mayor a 0,4 y significativo de S (Tabla 2). Las bacterias que cumplen con los requisitos mencionados anteriormente son 17. En estos organismos nueves codones: $GAC_{(Asp)}$, $GUU_{(Val)}$, $GCU_{(Ala)}$, $UCC_{(Ser)}$, $CAA_{(Gln)}$, $CGU_{(Arg)}$, $GAA_{(Glu)}$, $CAC_{(His)}$ and $GGU_{(Gly)}$, aparecen como codones óptimos altamente conservados ($\chi 2$, $p<0,05$, > 80% de los organismos) (Figura 8). Como es esperable, los cuatros codones óptimos universales, $UUC_{(Phe)}$, $UAC_{(Tyr)}$, $AUC_{(Ile)}$ and $AAC_{(Asn)}$, también aparecen como preferidos en forma significativa.

Por otra parte, se identificaron a 17 codones que son significativamente evitados por los genes de alta expresión: $UCA_{(Ser)}$, $UAU_{(Tyr)}$, $AAU_{(Asn)}$, $CCG_{(Pro)}$, $CAG_{(Gln)}$, $UUU_{(Phe)}$, $AGC_{(Ser)}$, $AUA_{(Ile)}$, $AGG_{(Lys)}$, $CGA_{(Arg)}$, $CGG_{(Arg)}$, $GCG_{(Ala)}$, $GAU_{(Asp)}$, $GGA_{(Gly)}$, $GGG_{(Gly)}$, $CAU_{(His)}$ and $GAG_{(Glu)}$ ($\chi 2$, $p < 0,05$, > 80% en los organismos) (Figura 8). El número de codones identificados como óptimos en cada genoma varía entre 15 y 21, y entre ellos la coincidencia varía de un 65% hasta el 89% con el ARNt isoaceptor presente en el genoma (Figura 8). Los codones no óptimos (rechazados) por los genes que codifican a las proteínas ribosomales presentan una leve tendencia a terminar con la base G.

No se identificaron ARNt isoaceptores que se aparean con enlaces del tipo Watson y Crick (WC) para los codones GUU (Val) y GGU (Gly). Los codones óptimos que no presentan genes de ARNt pueden ser reconocidos por tambaleo, por los anticodones GAC (codón GUC) y GCC (codón GGC) (Figura 8).

Una posible explicación por la elección de los codones GGU y GUU es para poder mantener el nivel de energía intermedia entre el enlace del codón con el anticodón (Grosjean y Fiers, 1982). Se han propuesto algunas explicaciones para el hecho de que el codón óptimo no es reconocido por el ARNt con más copias (Kahali et al. 2008), por ejemplo diferentes niveles de expresión que llevan a una concentración similar de ARNt independientemente de número de copias de genes, modificaciones post-transcripcionales en la primer posición del anticodón en algunos de los ARNt iso-aceptores. Por último, el que no haya ARNt isoaceptor mas abundante que se asocie con enlaces del tipo WC con respeto al codón óptimo, puede ser el resultado de la reducción en el efecto de la selección natural.

Como ya se dijo, se ha demostrado que los codones más frecuentemente utilizados por los genes de alta expresión, son aquellos que son reconocidos por el ARNt isoaceptor más abundante (Ikemura 1985; Kanaya et al. 2001; Kanaya et al. 1999). Sin embargo, se mostró que cambios en la concentración ya sea por mutaciones o pérdida en la cantidad de genes de ARNt no presentaron cambios significativos en la eficiencia de la elongación (Pop et al. 2014). De esta manera no queda claro cuáles son los beneficios a nivel de la eficacia en la traducción y elongación, la implicancia del sesgo en uso de codones sinónimos y la concentración de los ARNt.

En resumen, este resultado sugiere que el efecto del sesgo mutacional en el patrón observado es mínimo, y por tanto, la selección para la traducción podría estar operando, favoreciendo codones específicos en genes de ata expresión y en las regiones conservadas.

Tabla 2. Contenido de GC y GC3, coeficiente de Sharp y los coeficientes de correlación de los ejes generados mediante el WCA y las propiedades analizadas (MELP y distancia sinónima).

Micro-organism	Media GC ± SD[a]	Media GC3s ± SD[b]	MELP[c]	Var[d]	R(RCvsRG)[f]	S[g]	Random[h]	
Comamonas testosteroni CNB 2	0.62 ± 0.04	0.80 ± 0.01	-0.81	**23.5**	0.65	1.177	-0.274	0.177
Comamonas testosteroni ATCC 11996	0.62 ± 0.04	0.79 ± 0.10	-0.8	8.0	0.63	1.091	-0.385	0.179
Comamonas testosteroni S44	0.62 ± 0.04	0.79 ± 0.10	-0.78	7.7	0.66	1.063	-0.292	0.182
Comamonas testosteroni KF 1	0.62 ± 0.05	0.80 ± 0.11	0.82	7.5	0.65	0.983	-0.369	0.196
Ramlibacter tataouinensis TTB310	0.70 ± 0.04	0.94 ± 0.06	0.76	3.9	0.59	0.691	-0.832	0.465
Delftia acidovorans SPH 1	0.66 ± 0.05	0.88 ± 0.10	-0.76	6.0	0.36	0.604	-0.427	0.257
Delftia Cs1 4	0.67 ± 0.05	0.89 ± 0.09	-0.82	**31.7**	0.38	0.598	-0.462	0.257
Delftia sp. JD2	0.67 ± 0.05	0.88 ± 0.11	-0.58	5.9	0.43	0.595	-0.57	0.275
Verminephrobacter At4	0.66 ± 0.06	0.84 ± 0.13	-0.7	**29.2**	0.73	0.535	-0.469	0.29
Variovorax paradoxus B4	0.67 ± 0.04	0.89 ± 0.07	-0.84	4.8	0.33	0.534	-0.416	0.233
Acidovorax NO 1	0.65 ± 0.05	0.83 ± 0.09	0.81	6.3	0.36	0.508	-0.306	0.192
Variovorax paradoxus S110	0.68 ± 0.04	0.91 ± 0.07	0.85	5.1	0.31	0.508	-0.342	0.227
Acidovorax KKS102	0.65 ± 0.04	0.85 ± 0.08	-0.88	7.6	0.31	0.465	-0.398	0.205
Acidovorax avenae ATCC 19860	0.69 ± 0.04	0.91 ± 0.07	0.73	6.2	0.43	0.463	-0.772	0.346
Polaromonas CF318	0.65 ± 0.04	0.89 ± 0.08	-0.79	6.0	0.31	0.462	-0.325	0.192
Acidovorax citrulli AAC00 1	0.68 ± 0.04	0.90 ± 0.09	-0.77	5.2	0.47	0.441	-0.595	0.358
Acidovorax delafieldii 2	0.65 ± 0.05	0.83 ± 0.10	0.4	7.6	0.4	0.402	-0.369	0.214
Acidovorax radicis N35	0.65 ± 0.04	0.85 ± 0.08	-0.83	6.7	0.25	0.388	-0.392	0.219
Polaromonas JS666	0.62 ± 0.04	0.81 ± 0.10	0.66	4.7	0.37	0.386	-0.293	0.155
Verminephrobacter eiseniae EF01 2	0.65 ± 0.05	0.86 ± 0.10	-0.4	**24.9**	0.64	0.366	-0.362	0.271
Variovorax paradoxus EPS	0.67 ± 0.04	0.89 ± 0.07	-0.84	6.2	0.16	0.332	-0.461	0.271
Acidovorax CF316	0.68 ± 0.04	0.89 ± 0.08	0.37	7.4	0.26	0.31	-0.332	0.251
Variovorax CF313	0.67 ± 0.04	0.89 ± 0.07	-0.81	5.5	0.17	0.304	-0.401	0.262
Hylemonella gracilis ATCC 19624	0.65 ± 0.04	0.86 ± 0.08	-0.42	**22.0**	0.65	0.301	-0.368	0.226
Polaromonas naphthalenivorans CJ2	0.62 ± 0.06	0.82 ± 0.13	0.71	4.7	0.09	0.288	-0.368	0.205
Acidovorax JS42	0.66 ± 0.04	0.87 ± 0.09	-0.77	5.3	0.34	0.269	-0.448	0.265
Acidovorax ebreus TPSY	0.67 ± 0.04	0.88 ± 0.07	-0.83	6.3	0.22	0.147	-0.422	0.232
Alicycliphilus denitrificans K601	0.68 ± 0.05	0.91 ± 0.08	-0.81	4.7	0.22	0.143	-0.482	0.344
Alicycliphilus denitrificans BC	0.68 ± 0.05	0.91 ± 0.09	-0.77	4.7	0.18	0.13	-0.601	0.369
Rhodoferax ferrireducens T118	0.60 ± 0.04	0.77 ± 0.10	-0.47	**21.6**	0.05	0.034	-0.204	0.144
Hydrogenophaga PBC	0.68 ± 0.04	0.91 ± 0.07	0.72	5.5	0.21	-0.018	-0.497	0.298
Comamonadaceae bacterium CR	0.58 ± 0.06	0.71 ± 0.10	0.88	**25.5**	-0.48	-0.203	-0.401	0.25

[a]El contenido de %G+C del genoma.

[b]El contenido de %G+C el genoma en la tercera posición del codón.
[c]Coeficiente de correlación de Pearson's (r) de las posiciones de los genes en el eje de expresión del WCA contra los valores de respectivo de MELP, en rojo y amarillo se representan la correlación con el seguno y el primer eje respectivamente.
[d] Se muestra la variabilidad explicada por el primer y el segundo eje genereado por el WCA..
[f] Correlación entre la frecuencia en el uso de los codones sinónimos entre las regiones conservadas en genes ortólogos con respecto a las proteínas ribosomales.
[g] La fuerza de selección del sesgo codón de uso seleccionado (S) calculada de los genes de alta expresión.
[h] Se presenta el rango del 95% de los valores de S entre los 1000 genes elegidos al azar.

Figura 8. Codones preferidos y evitados por los GAEs y por las regiones conservadas de las proteínas ortólogas para cada organismo. Se utilizó el estadístico $\chi 2$ de contingencia para poder determinar que codón es preferido o evitado en forma significativa por los genes GEA y las regiones conservas con respecto a los genes GBE y las regiones no conservas ($p < 0.05$). En rojo están representados los codones preferidos por los GEA ($p<0.05$), mientras que en azul esta los codones evitados ($p< 0.05$). Se indica el número de copias de genes ARNt para cada organismo.

4.5 Estimación de la distancia molecular

Siguiendo al trabajo de Sharp y Li (1987), se investigó el efecto de la selección negativa (purificadora) en los sitios sinónimos estudiando la tasa de sustitución sinónima (dS) en los genes ortólogos. Se ha demostrado que cuando existe selección a nivel de traducción, la tasa de sustitución sinónima (dS) y el MELP (índice de expresión) presentan una correlación negativa (Sharp y Li, 1987). De esta forma los genes que son considerados de alta expresión van a presentar una menor tasa de cambios sinónimos en comparación con los genes de baja expresión, si la selección favorece la misma elección del uso de codones sinónimos en los genes de alta expresión, desde el último ancestro común.

A partir del resultado de la correlación entre el MELP y el dS entre todos los organismos en estudio, se puede inferir en forma general que los organismos que presentan un valor de S significativo y mayor a 0,4 presentan una correlación negativa y significativa (Figura A4).

Sin embargo es importante de destacar también que las bacterias *C. testosteroni* CNB y *C. testosteroni* S44, se caracterizan por presentar valores altos de S, sin embargo presentan un valor de correlación bajo (Figura 9). Este patrón puede ser fácilmente explicado como consecuencia de la distancia filogenética entre estas dos bacterias: debido al corto tiempo de divergencia, los genes ortólogos no han podido acumular suficientes cambios sinónimos, para generar el patrón esperado.

Por otro lado la bacteria *Ramlibacter tataouinensis* TTB310 presenta un valor alto de S (0,69), y cuando se le compara el MELP con el dS contra los organismos del género *Delftia* y *Comamonas* presentan una correlación negativa pero baja. Este comportamiento se puede asociar con cambios en la elección de los codones óptimos entre los organismos mencionados anteriormente. Estos cambios podrían ser el resultado de la selección divergente que actúa en cada linaje específico, dirigiendo la acumulación de más cambios sinónimos en los genes de alta expresión (Figura 9). La posición de *R. tataouinensis* TTB310 en la filogenia y la identificación de los codones óptimos apoyan esta explicación.

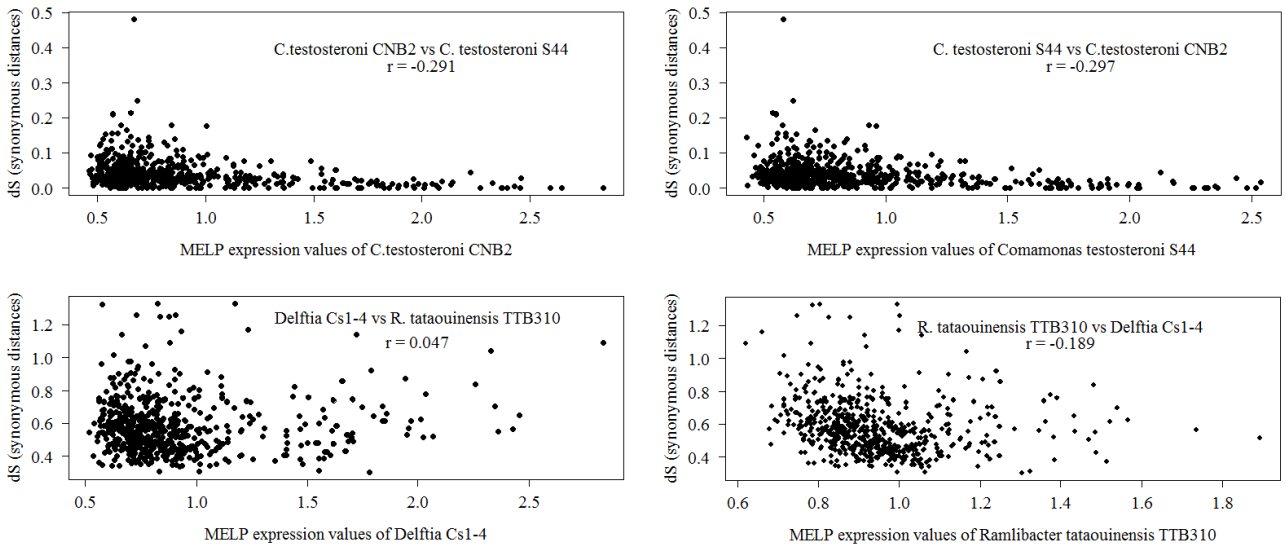

Figura 9. Se muestra la regresión de los valores de MELP contra los valores estimados de dS.

Para poder testear si existe selección a nivel de la fidelidad se correlacionó los valores de dS y dN de cada bacteria. Se generaron tres grupos para realizar este análisis. El primer grupo (Selección fuerte) en el cual están incluidos los organismos que presentan un valor de S mayor a 0,4 y significativos. El segundo grupo (selección moderada) son los organismos que presentan valor de S significativos pero sin superar el valor de 0,4 y por último el tercer grupos (selección débil) que presentan valores de S no significativos y que no superan el valor 0,4. A partir de las correlaciones se puede inferir que las bacterias que presenta selección fuerte o moderada presentan una alta fidelidad o sea que los genes que presentan menos cambio no sinónimos, también presentan a su vez menos cambios sinónimos. Sin embargo los organismos del tercer grupo (débil) los valores de las correlaciones son bajos (Figura 10, Figura A5). De esta manera la conservación aminoacídica se ve reflejada a nivel de las posiciones sinónimas en los codones.

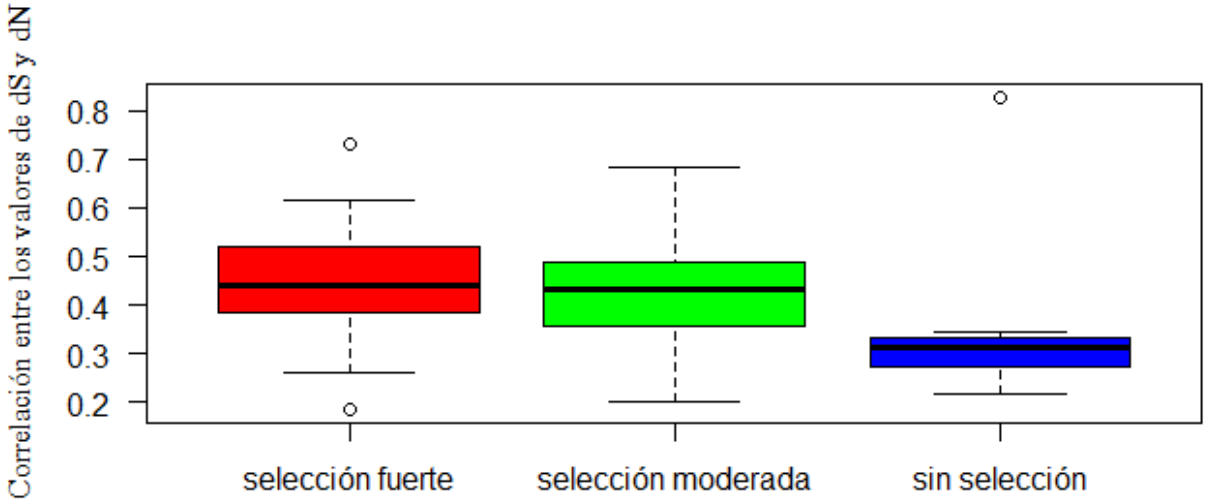

Figura 10. Boxplot de los valores de correlación entre el dS y dN. Se muestran a las bacterias que presentan selección o sea un valor S significativo y mayor a 0.4, las bacterias que presentan una selección moderada tiene un valor de S significativo pero no supera el valor de 0.4. Por últimos las bacterias que no presentan selección tienen un valor de S no significativo e inferior a 0.4.

4.6 **Análisis de la energía mínima libres de plegamiento del ARNm en relación con el sesgo en el uso de codones.**

Se comparó la energía mínima libre normalizada (MELN) entre los genes GAE (Proteínas ribosomales) y GBE (resto de los genes), con el fin de buscar posibles asociaciones. Se ha sugerido que las secuencias de ARNm presentan un mayor valor negativo de energía libre de plegamiento que un conjunto de secuencias de ARNm aleatorios (Seffens y Digby 1999). Este comportamiento puede reflejar una selección a nivel del plegamiento del ARNm. Hay evidencias que muestran que los codones óptimos se encuentran disminuidos en los sitios pareados en comparación con respecto a los sitios no pareados del ARNm (Stoletzki 2008). Alto valores de GC conducen generalmente a estructuras más estables del ARNm y este factor puede interferir con la iniciación de la traducción. Estructuras menos estables a nivel 5' del ARNm se correlacionan con los niveles de expresión más altos (Kudla et al. 2009). A su vez el contenido de GC es una de las principales variables que afectan al uso de codones sinónimos (Chen et al. 2004). Los genes de las proteínas ribosomales presentan un menor contenido de GC y MELN con respecto al resto de los genes del genoma, y esta diferencia es significativa (W, $p<0,05$) (Tabla 3). Como las especies en estudio están relacionadas por una historia en común los datos de cada especie puede que no sean independientes estadísticamente. Para poder corregir este efecto se analizó de nuevo utilizando el método de contraste independiente filogenético (PIC) (Ver Materiales y Métodos) (Felsenstein 1985). Se van a

mostrar y a discutir los resultados sin las correcciones de PIC, a excepción de que los resultados corregidos por PIC presentan un comportamiento distinto al de los resultados no corregidos.

Cuando se grafica la diferencia de GC entre los GAE y los GBE y se correlaciona con el GC genómico, se puede apreciar una correlación de -0.41 (p=0.02) (Figura 11). De este análisis se puede inferir que los organismos que presentan una mayor diferencia en el contenido de GC entre los GAE con respeto a los GBE, se asocian débilmente con los genomas más ricos en GC. En otras palabras, la diferencias de GC entre los GAE y los GBE no presentan un gran influencia por el contenido de GC genómico. Sin embargo, cuando se realiza la correlación entre las diferencias de MELN entre los GAE y GBE con respeto al contenido de GC genómico, se puede ver una correlación fuerte con una valor de r de 0,74 (p=1x10e-6) (Figura 11). De esta manera se podría afirmar que la diferencias de MELN entre los GAE y los GBE pueden estar influenciados por el contenido de GC genómico. Los organismos más ricos en el contenido de GC presentan una mayor diferencia entre la MELN de los GAE con respecto a los genes GBE. Este patrón podría ser la consecuencia de la selección natural operando a nivel de la estructura en los ARNm.

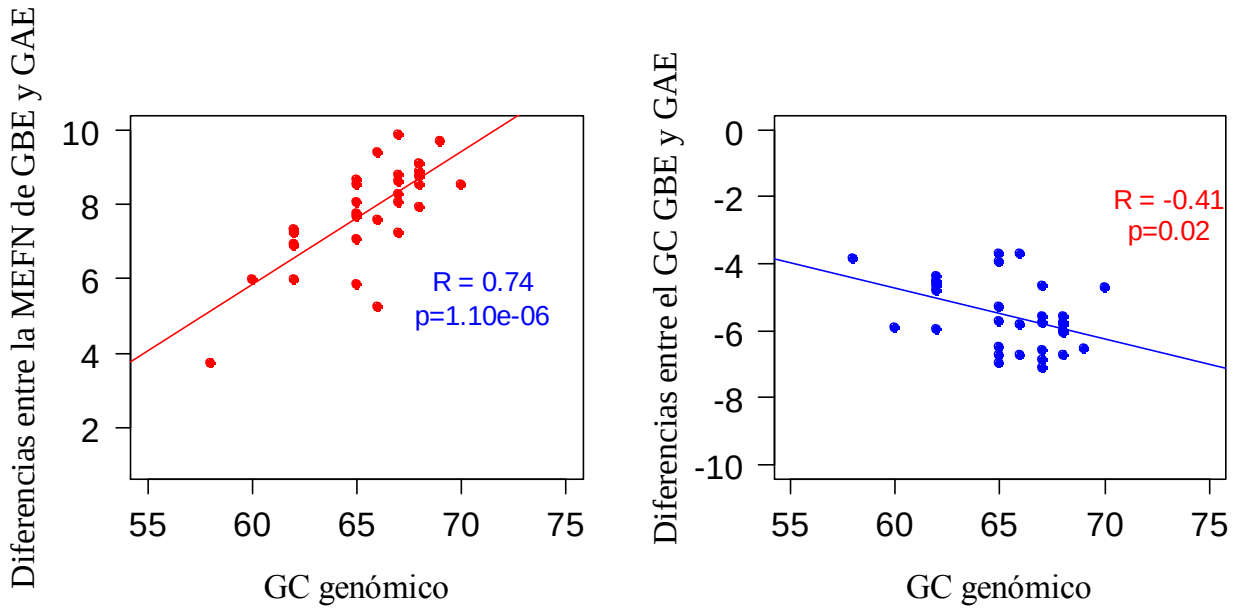

Figura11. Se muestra la regresión entre los valores de GC genómicos en la diferencias de Mínima energía de plegamiento normalizada y la diferencias de GC entre los genes de lata expresión con respecto a los genes de baja expresión.

Con el fin de poder identificar cuál es el grupo de secuencias (GAE o GBE) responsable del patrón mencionado anteriormente, se comparó el contenido de GC y la MELN de los GAE y GBE con respecto al contenido de GC genómico. Del resultado de la comparación del GC de los GAE y GBE con el GC genómico se puede apreciar que presentan un comportamiento similar con pendiente casi idéntica con valores de 1,01 y 0,99 respectivamente, a pesar de que presentan valores promedios

distintos de GC (Figura 12).

Por otra parte, cuando se analiza el comportamiento de los valores de la MELN con respecto el contenido de GC de los genomas, se puede apreciar que tienden a presentar el mismo comportamiento pero con diferentes pendientes. Los valores de las pendientes para los GAE y GBE son -1,34 y -1,03, respectivamente (Figura 12). A partir de este resultado, se puede inferir que las proteínas ribosomales presentan un valor de MELN diferente de lo esperado por el contenido de GC en los organismos más ricos en GC. Es sabido que la MELN es proporcional al contenido de GC, por lo tanto se esperaría un comportamiento similar al presentando por el contenido de GC de los GAE y GBE (Seffens y Digby 1999). Sin embargo se ha demostrado que los genes más expresados presentan una mayor estructura secundaria en comparación con los menos expresados (Del Campo et al. 2015; Gorochowski et al. 2015; Park et al. 2013). Asimismo, se ha demostrado que las estructura secundaria de los ARNm de los genes "esenciales" tienden a ser más conservadas en comparación con los genes "no esenciales" (Chursov et al. 2013). De esta manera, se puede pensar que el ARNm no sólo lleva la información de los aminoácidos a codificar sino que el mismo podría estar implicado en forma directa o indirecta con el correcto plegamiento de las proteínas a sintetizar (Chursov et al. 2013; Faure et al. 2016). Sin embargo, es sorprendente que las proteínas ribosomales de los genomas ricos en GC tiendan mantener un MELN menor de lo esperado por el contenido de GC.

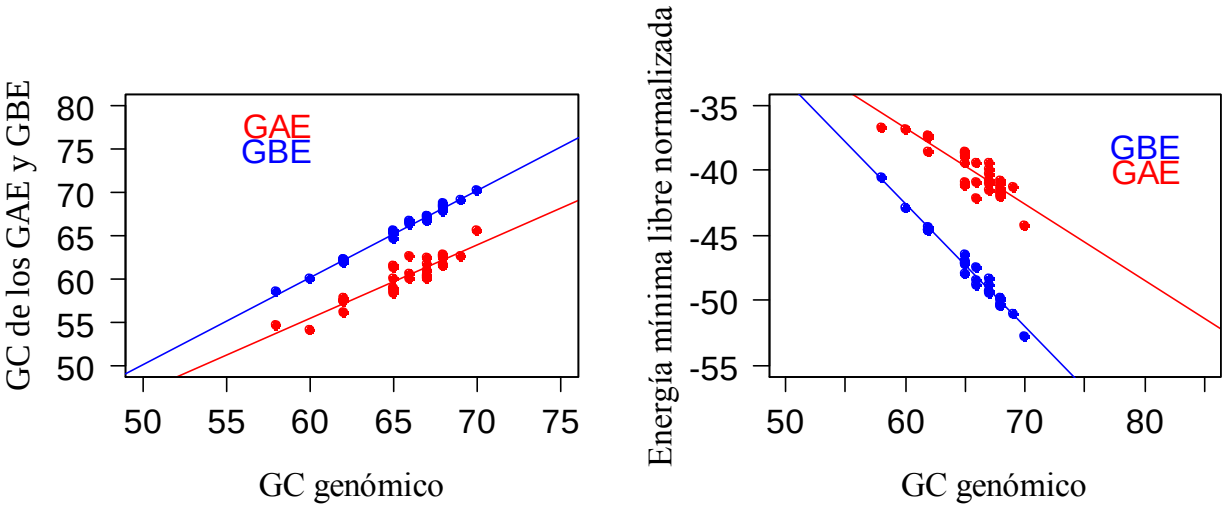

Figura 12. Se muestra la regresión entre le contendeos de GC genómico con respeto al contenido de GC y Energía mínima de libre normalizada de los genes de alta y baja expresión.

Por lo tanto, las proteínas ribosomales son menos estables con respecto a los GBE, de esta manera se esperaría que presenten menos estructuras potenciales en comparación con los GBE. A su vez, estructuras estrechamente plegadas en el ARNm llevan a una dificultad en la iniciación de la

traducción y por lo tanto se disminuye la síntesis de proteínas (Kudla et al. 2009; Pop et al. 2014). Sin embargo, se ha demostrado que la gran mayoría de las estructuras en el ARNm parecen no afectar el recorrido del ribosoma a través del ARNm (Del Campo et al. 2015). El sesgo en la selección de codones favorece una estructura del ARNm que contribuye a la estabilidad de plegado (Seffens y Digby, 1999). Esto sugiere que la estructura del ARNm puede dar forma a los niveles de expresión (Duan et al. 2003). Este patrón parece ser más pronunciado en los genomas ricos en GC. Una posible estrategia de los ARNm de las proteínas ribosomales para poder ser traducidos en forma más rápida sería presentando una menor cantidad de GC y MELN con respecto al resto de los genes, en estos organismos.

Tabla 3. Medias en el contenido de GC y energía mínima libre normalizada.

Micro-organism	**Media GC[A]**	**Media GC[B]**	**Media MFEN[C]**	**Media MFEN[D]**	**Medias al azar MFEN[E]**	
Acidovorax avenae ATCC 19860	62.5	69.06	-41.3	-50.99	-54.018	-49.376
Acidovorax citrulli AAC00 1	62.63	68.63	-41.42	-50.49	-53.439	-48.828
Acidovorax ebreus TPSY	60.35	67.22	-40.81	-49.39	-52.071	-47.819
Acidovorax JS42	60.61	66.44	-40.95	-48.52	-51.393	-46.915
Acidovorax KKS102	58.88	65.35	-38.75	-47.25	-50.083	-45.803
Alicycliphilus denitrificans BC	62.45	68.26	-41.64	-50.39	-53.851	-48.672
Alicycliphilus denitrificans K601	62.67	68.25	-41.85	-50.36	-53.644	-48.684
Comamonadaceae bacterium CR	54.68	58.53	-36.8	-40.54	-43.64	-38.824
Comamonas testosteroni CNB 2	57.42	62.11	-37.49	-44.7	-47.265	-43.274
Delftia acidovorans SPH 1	59.94	66.68	-39.46	-48.83	-51.891	-47.137
Delftia Cs1 4	60	67.09	-39.44	-49.3	-52.431	-47.613
Polaromonas JS666	57.81	62.17	-38.66	-44.63	-47.275	-43.221
Polaromonas naphthalenivorans CJ2	56.14	62.12	-37.39	-44.7	-48.112	-42.987
Ramlibacter tataouinensis TTB310	65.47	70.21	-44.3	-52.81	-55.717	-51.223
Rhodoferax ferrireducens T118	54	59.91	-36.92	-42.87	-45.724	-41.537
Variovorax paradoxus B4	61.65	67.24	-41.03	-49.3	-51.895	-47.889
Variovorax paradoxus EPS	60.93	66.71	-40.33	-48.35	-50.924	-46.977
Variovorax paradoxus S110	61.65	67.71	-40.77	-49.85	-52.574	-48.367
Verminephrobacter eiseniae EF01 2	61.41	65.13	-40.9	-47.94	-51.006	-46.325
Acidovorax CF316	61.38	68.11	-41.1	-49.94	-52.759	-48.391
Acidovorax delafieldii 2AN	58.78	65.49	-38.95	-47	-49.872	-45.46
Acidovorax NO 1	58.91	64.64	-38.78	-46.54	-49.576	-44.974
Acidovorax radicis N35	58.3	65.29	-38.56	-47.21	-50.044	-45.714
Comamonas testosteroni ATCC 11996	57.43	61.97	-37.52	-44.45	-47.115	-42.974
Comamonas testosteroni KF 1	57.45	62.25	-37.37	-44.68	-47.582	-43.13
Comamonas testosteroni S44	57.34	61.92	-37.5	-44.37	-47.332	-42.904
Hydrogenophaga PBC	62.7	68.47	-42.04	-49.96	-52.466	-48.483
Hylemonella gracilis ATCC 19624	61.32	65.29	-41.17	-47.01	-49.706	-45.63
Polaromonas CF318	60.01	65.3	-39.47	-47.14	-50.075	-45.67
Variovorax CF313	62.36	67.02	-41.57	-48.78	-51.402	-47.365
Verminephrobacter At4	62.59	66.28	-42.23	-47.46	-50.914	-45.662
Delftia sp. JD2	60.13	66.7	-39.99	-48.77	-52.047	-46.944

[a] EL contenido de GC de los genes de las proteínas ribosomales.
[b] EL contenido de GC de todos los genes sin los genes de las proteínas ribosomales
[c] Mínima energía libre normalizada de los genes de las proteínas ribosomales.
[d] Mínima energía libre normalizada de todos los genes sin los genes de las proteínas ribosomales .

El resultado obtenido podría estar asociado con la función de estos genes que se consideran de alta expresión. Se estima que aproximadamente el 60% de todas las actividades de transcripción celular

están vinculadas a la síntesis de las proteínas ribosomales en una célula en crecimiento rápido (Warner 1999) y 40% de la energía total de una célula de *E. coli* es dirigida hacia la síntesis de proteínas ribosomales (Wilson i Nierhaus 2007). Se sabe que la elección de los codones preferidos varía entre los organismos (Sharp y Li 1986). A partir de nuestros resultados se propone que una estrategia para codificar de forma rápida y eficiente las proteínas ribosomales, es que tengan un bajo valor de MELN en comparación con los otros genes, de esta manera la maquinaria de traducción puede tener una mayor accesibilidad al ARNm. Por lo tanto un factor que podría estar determinando la opción de los codones óptimos en los genes de alta expresión sea la estructura secundaria del ARNm.

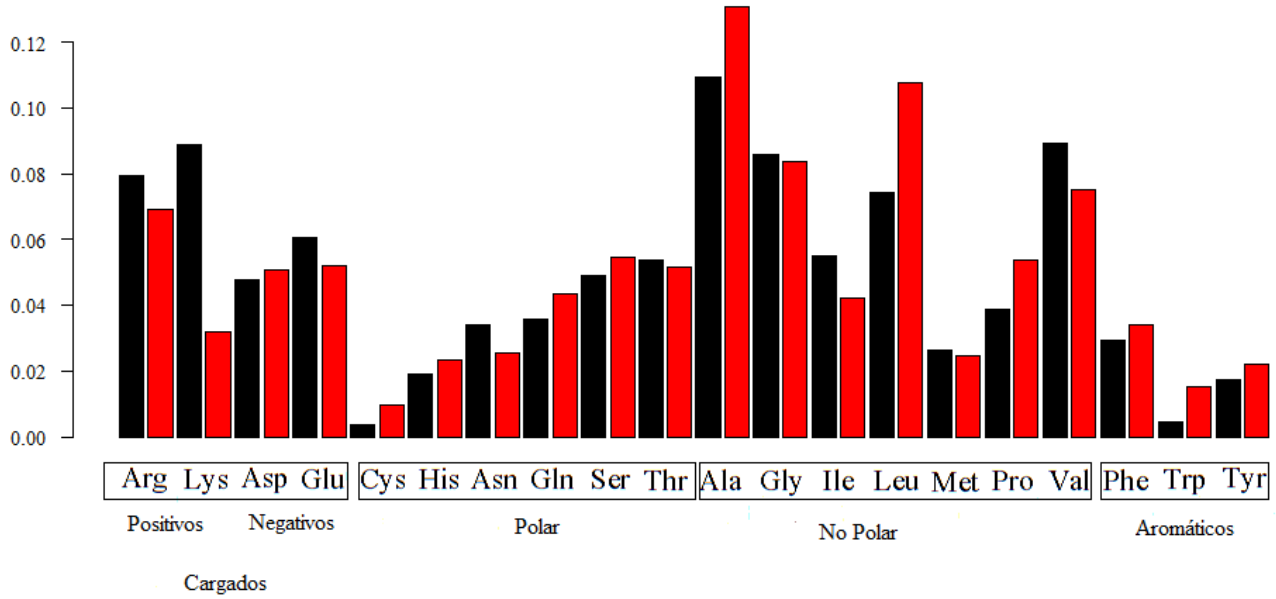

Figura 13. Uso de aminoácidos de las proteínas ribosomales (Negro) y las proteínas no ribosomales (Rojos).

Por otra parte, el uso de aminoácidos en las proteínas ribosomales podría estar condicionando el uso de codones sinónimos y de esta forma la proporción de bases en el gen, y por último condicionando la MELN.

Las proteínas ribosomales presentan una mayor proporción de aminoácidos con carga positiva (Lys y Arg) en comparación con el resto de las proteínas. Los aminoácidos cargados negativamente a pH fisiológico (Asp y Glu) presentan comportamientos opuestos: mientras que Glu se encuentra en mayor proporción en las proteínas ribosomales, Asp es más frecuente en las no ribosomales (Figura 13, Tabla A3).

Las sumas de las medias de los aminoácidos con carga positiva (Arg, Lys) en las proteínas ribosomales se encuentran en mayor proporción en comparación con los aminoácidos con carga negativa (Asp, Glu), este resultado concuerda en forma general con los resultados de otro autor

(Lott et al. 2013). La mayor proporción de aminoácidos con cargas positivas puede estar implicado en la interacion entre las proteínas ribosomales con el ARNt, ya que la interacción entre estas dos moléculas presenta un importante componente electrostático (Trylska et al. 2004). Por otro lado, la suma de las medias de los aminoácidos positivos (Arg, Lys) presenta el mismo valor que la suma de las medias de los aminoácidos negativos (Asp, Glu), en las proteínas no ribosomales. Para poder ver la magnitud y dirección de las preferencias de los aminoácidos, se analizaron las diferencias entre los dos grupos de secuencias. Encontramos que los aminoácidos que son utilizados en mayor proporción por las proteínas ribosomales son Ile, Met, Val, Thr, Asn, Lys, Glu, Arg y Gly (Figura 14, Tabla A3).

De esta manera, es probable que el uso de aminoácidos de las proteínas ribosomales tiende a disminuir el contenido de GC total y en todas las posisiciones de los codones (Figura 15). Cuando se analizan por separados a las bases en las tercera posición de los codones de los cuartetos naturales y de los cuartetos de los sextetos las bases C3 y G3 presenta diferencias significativas; sin embargo, G3 presenta una menor proporción en las proteínas ribosomales en comparación con C3 (Figura 16). De esta manera se podría inferir que a pesar de que el contenido de GC está condicionado por el sesgo en el uso de aminoácidos, éste tiende a diminuir más el contenido de GC evitando a los codones sinónimos terminados en G. Como se ha demostrado anteriormente, los codones óptimos no presentan ningún codon terminado en G, mientras que los codones rechazados por los genes de las proteínas ribosomales son en su mayoría terminados en G (Figura 8). Sin embargo, se ha demostrado que el uso relativo celular de los aminoácidos está asociado al contenido de GC de los organismos (Moura et al. 2013). El contenido de GC "dirige" al uso de codones y éste último afecta, de manera indirecta, al uso de aminoácidos ya que hay cambios en la tercera posición de los codones (y algunos en la primera) que no son sinónimos, y todos los que ocurren en la posición dos son no sinónimos (Knight et al. 2001). Es de destacar que es difícil poder analizar en forma separada las fuerzas evolutivas que están determinando el sesgo en el uso de aminoácidos, la energía mínima de plegamiento, modulando y cuantificando su efecto en la eficacia de las tasa y la eficacia de la traduccion (Shabalina et al. 2013). Recientemente se encontró una posible asociación entre las regiones estables del ARNm con regiones de proteínas que codifican dominios compactos y proteínas de gran tamaño. De esta manera las estructuras secuandarias del ARNm, podrían estar implicadas como posibles indicadores co-traduccionales del plegamiento de las proteínas (Faure et al. 2016).

De esta forma, se propone que el sesgo en el uso de aminoácidos, el sesgo en el uso de codones y la energía mínima de plegamiento deben de tener una participacion significativa en la estrategía para optimizar la transcripción y traducción en los organismos en estudio.

No se puede inferir que este comportamiento sea general en todos los procariotas, ya que se ha demostrado que éstos organismos presentan diferentes intensidades de fuerzas selectivas (Ran et al. 2014). Se propone que el uso de codones sinónimos puede estar reflejando una posible estrategia en la que la MELN de los genes de las proteínas ribosomales, module la síntesis de estas proteínas evitando potenciales estructuras secundarias.

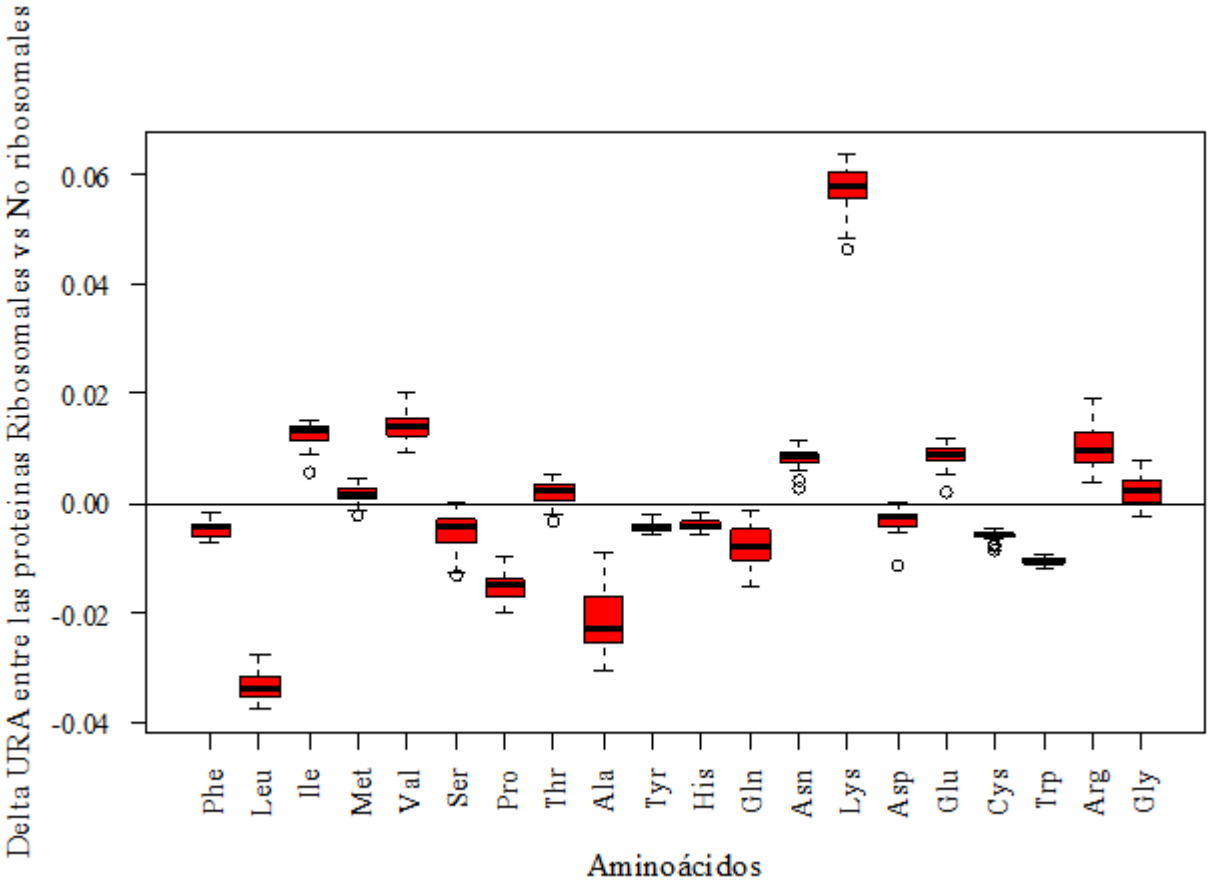

Figura 14. Diferencias entre en el uso de aminoácidos de las proteínas ribosomales con respecto al resto de los genes. Todas las diferencias son estadísticamente significativas (W, $p<0.05$), a excepción de Cys (Tabla A4).

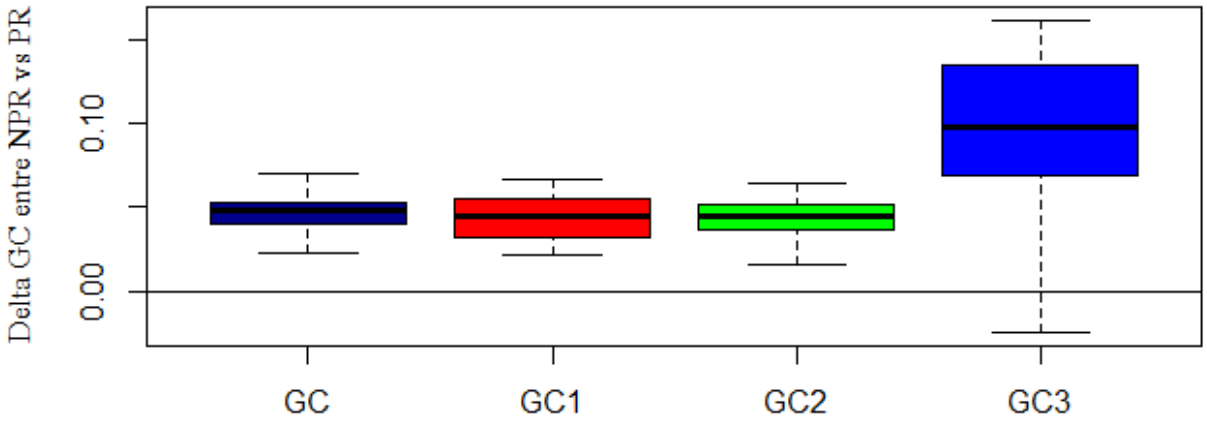

Figura 15. Diferencias en el contenido de GC total y en la primera, segunda y tercera posición de los codones de los genes de las proteínas ribosomales con respecto al resto de los genes de genoma en los 32 organismos en estudio. El contenido de GC3 se calculó a partir de las secuencias tomando solamente a los cuartetos naturales y a los pertenecientes de los sextetos. Todas las diferencias son estadísticamente significativas. (W, $p<0.05$) (Tabla A3).

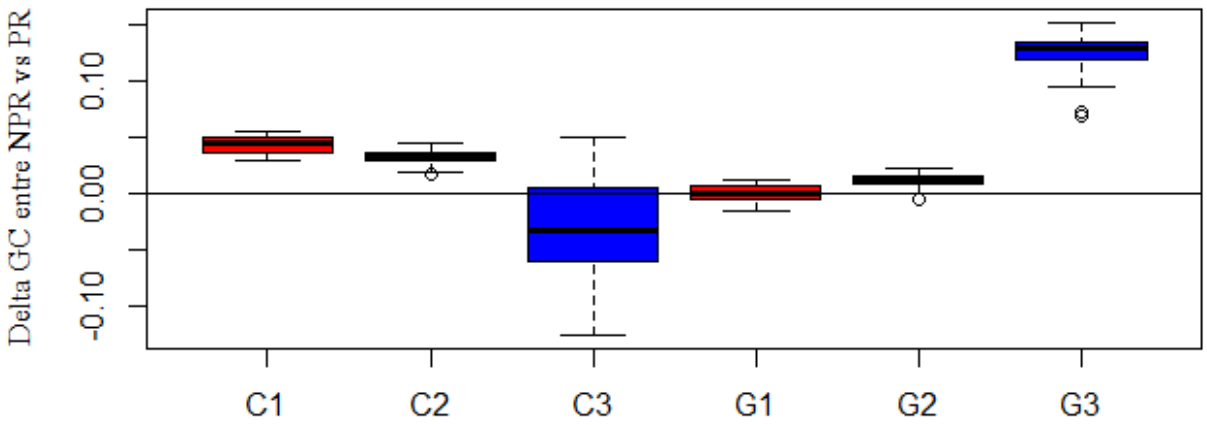

Figura 16. Diferencias en el contenido de G y C en la primera, segunda y tercera posición de los codones de los genes de las proteínas ribosomales con respecto al resto de los genes de genoma en los 32 organismos. El contenido de C3 y G3 se calculó a partir de las secuencias tomando solamente a los cuartetos naturales y a los pertenecientes de los sextetos. Todas las diferencias son estadísticamente significativas (W, $p<0.05$), a excepción de G1 (Tabla A3).

4.8 Uso de aminoácidos entre las proteínas de membrana, proteínas citoplasmáticas y entre las regiones TM y NTM

Las proteínas de membrana presentan en promedio un contenido más elevado de los aminoácidos Ile, Val, Leu, Phe, Met, Gly, Ser y Trp en relación a las proteínas citoplasmáticas (Figura 17), mientras que sucede lo inverso con Cys, Tyr, Pro, His, Asp, Gln, Glu, Lys y Arg, el aminoácido Thr no presenta diferencia significativa (Figura 17, Tabla A4). Cys, a pesar de ser hidrofóbico, se encuentra en mayor cantidad en las proteínas citoplasmáticas, mientras que los aminoácidos hidrofílicos Gly y Ser se encuentran en mayor cantidad en las proteínas de membrana que en las proteínas citoplasmáticas.

La distribución de aminoácidos entre las regiones TM y las regiones NTM es, como se esperaba, diferente: las regiones NTM presentan en promedio un enriquecimiento de los aminoácidos hidrofílicos Thr, Ser, Pro, His, Asn, Asp, Gln, Glu, Lys, y Arg con respecto a las regiones TM, mientras que las TM tienden a enriquecerse en los aminoácidos hidrofóbicos Ile, Val, Leu, Phe, Met, Ala, Trp y con los aminoácidos hidrofílicos Gly y Tyr (Figura 17, Tabla A5). Las regiones TM están formadas por alfa hélices las que son estabilizadas mediante la intercalación de aminoácidos hidrofílicos e hidrofóbicos (Khrustalev y Barkovsky 2012; Wolfenden et al. 1979). Como es de esperar, los residuos hidrofóbicos son utilizados en mayor medida en las regiones TM, mientras que los hidrofílicos son utilizados mayormente en las regiones NTM (Figura 18, Tabla A5). El aminoácido Cys no presenta diferencia significativa en su utilización entre las dos regiones.

Las regiones TM presentan una correlación positiva y significativa entre el uso relativo de aminoácidos y los valores de hidrofobicidad de la escala Kyte-Doolittle (Figura 19). Por otra parte las regiones NTM, las proteínas citoplasmáticas y las proteínas de membrana no presentaron una correlación con respecto a los valores de hidrofobicidad de la escala Kyte-Doolittle (Figura 19). De esta manera se asegura de que las secuencias que clasificamos como TM, sean efectivamente tales.

Aminoácidos químicamente similares tienden a presentar la misma base en la segunda posición del codón, hecho que podría reflejar, en cierta forma el origen del código genético, ya que se ha propuesto que este se ha organizado a partir de las estructuras secundarias de las proteínas (Chiusano et al. 2000). Es de destacar que los aminoácidos que son preferidos por las regiones TM son hidrofóbicos y estos presentan codones que en su mayoría presentan en la segunda posición la base U, por lo que es de esperar que presenten una menor cantidad de GC en sus codones.

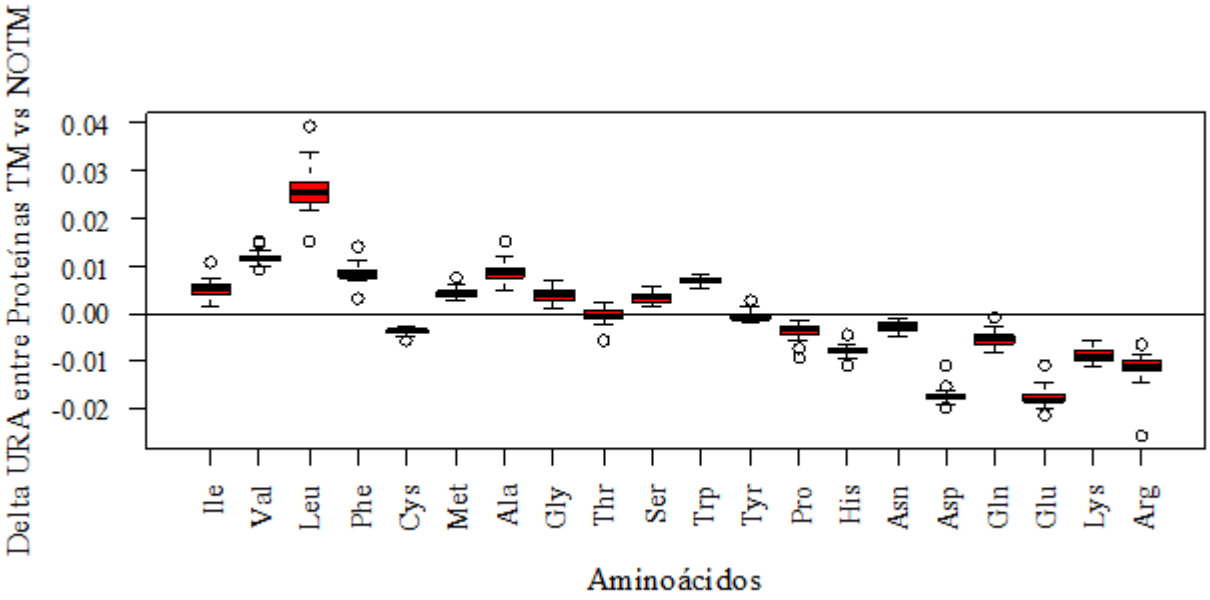

Figura 17. Diferencia en el uso de aminoácidos de las proteínas de membrana respecto a las proteínas citoplasmáticas. Los aminoácidos están ordenados por la escala de hidrofobicidad Kyte-Doolittle. Todas las diferencias son estadísticamente significativas (W, $p<0.05$), a excepción de Thr (Tabla A5).

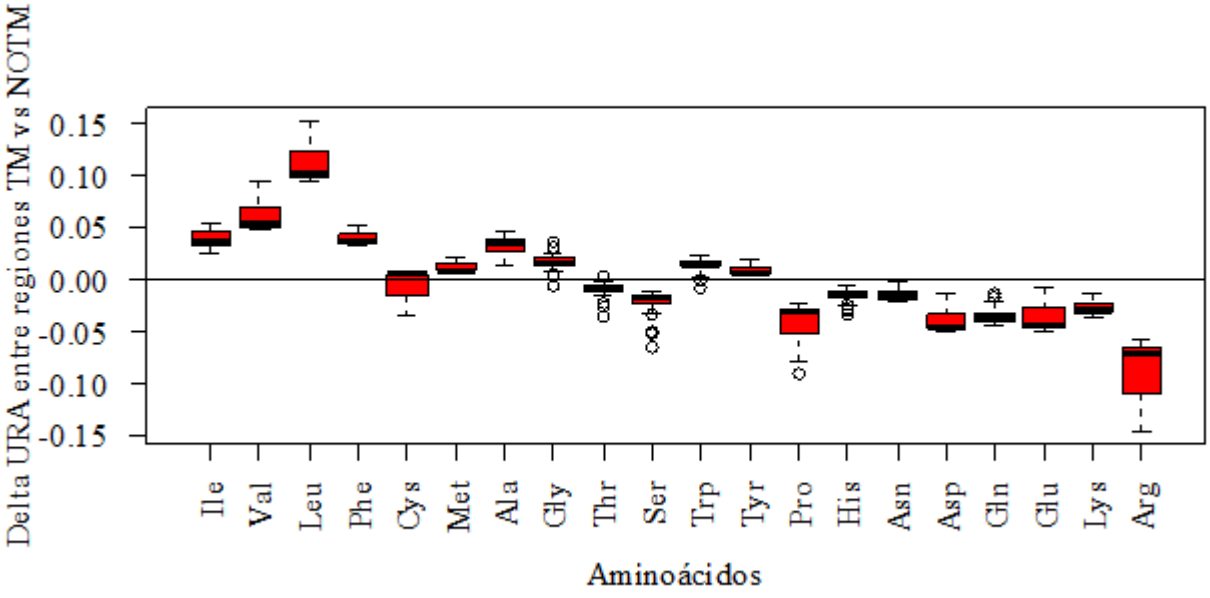

Figura 18. Diferencia en el uso de aminoácidos de las regiones TM respecto a las NTM. Los aminoácidos están ordenados por la escala de hidrofobicidad Kyte-Doolittle. Todas las diferencias son estadísticamente significativas (W, $p<0.05$), a excepción de Cys (Tabla A5).

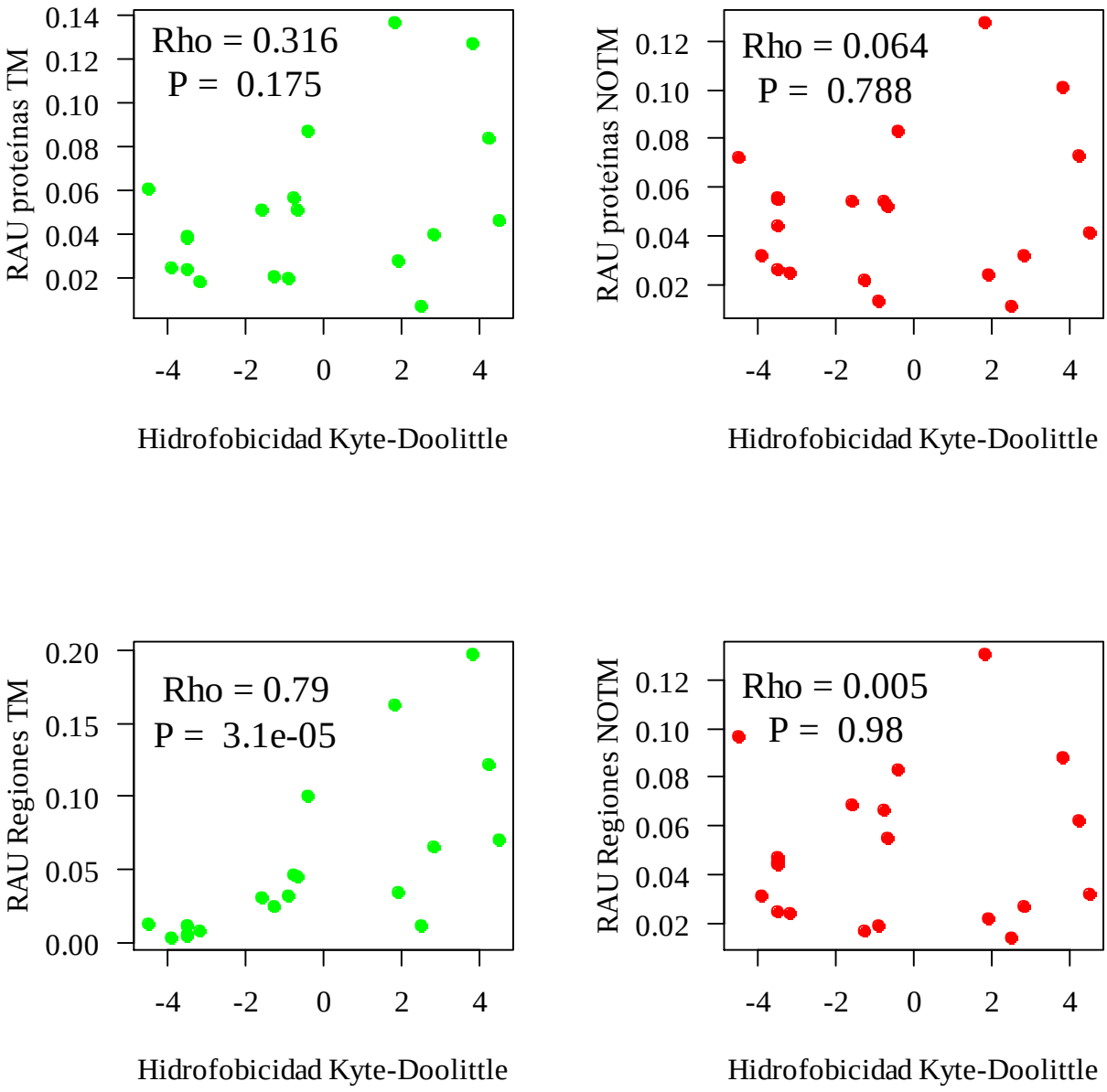

Figura 19. Correlación entre el uso de aminoácidos de las proteínas de membrana y proteínas no de membrana, regiones transmembrana y no transmembrana.

4.8.1 Análisis de usos de codones sinónimos entre las regiones transmembranas y no transmembranas de las proteínas de membrana.

Por definición las proteínas de membrana contienen regiones hidrofóbicas e hidrofílicas. Esto nos da la oportunidad de estudiar el UCS entre estas regiones en un mismo ARNm, ya que estas regiones son traducidas por la misma maquinaria de traducción, influenciadas por el mismo sesgo

mutacional y coexisten en el mismo ambiente celular. Con el fin de identificar a los codones preferidos por las regiones TM, se los comparó con los codones de las regiones NTM. Se observó una tendencia a que los codones CUG (Leu), UCG (Ser), GCG (Ala), GUG (Val), ACG (Thr), CCG (Pro) y UAU (Tyr) son utilizados en una mayor frecuencias en las regiones TM con respecto a las regiones NTM (Figura 23); y esta diferencia es estadísticamente significativa ($\chi 2$, $p < 0,05$). Los codones CUG, UCG, GCG, GUG, ACG, CCG y UAU presentan diferencias significativas ($\chi 2$, $p < 0,05$) en 100%, 100%, 80%, 80%, 80%, 100% y 60% de los organismos respectivamente (Figura 23). De estos codones, dos pertenecen a los "cuartetos de los sextetos", mientras que el resto pertenece a cuartetos con la excepción de UAU que pertenece a un dueto. Los codones preferidos por las regiones TM son en su mayoría terminados en G. Los codones CUG y GUG presentan al ARNt iso-aceptor que se enlaza en forma WC, mientras que por otro lado los ARNt iso-aceptores de los codones UCG, GCG, ACG, CCG y UAU están en minoría o no se ha identificado a los genes que codifican al ARNt que se enlace en forma de WC (Figura 23, Figura 8).

Los codones preferidos por las regiones TM con respecto a las regiones NTM, son en forma general evitados por los genes que codifican a las proteínas ribosomales. Como ya se ha mencionado anteriormente los codones preferidos por los genes que codifican a las proteínas ribosomales, son considerados óptimos, dándole fidelidad y velocidad a la traducción (Akashi 1994; Tuller et al. 2010). Existen numerosos ejemplos en los cuales el plegamiento y el nivel de expresión de las proteínas de membrana son afectadas por la utilización de codones sinónimos (Cortazzo et al. 2002; Makino et al. 1997). Según varios autores los cambios en los codones sinónimos generan un aumento o disminución en la velocidad de lectura del ARNm por el ribosoma, afectando la velocidad de síntesis proteica, influyendo entonces en el plegamiento de las proteínas nacientes (Cortazzo et al. 2002; Makino et al. 1997; Weygand-Durasevic e Ibba 2010). Esta es una de las razones por la cual las proteínas de membrana al traducirse a una velocidad no adecuada, se pliegan en forma incorrecta, exponiendo regiones hidrofóbicas llevando a la formación de agregados que son tóxicos (Cortazzo et al. 2002). Otro aspecto es el efecto del UCS sobre la estructura secundaria del ARNm, alterando la unión del ribosoma con el mensajero. Por ejemplo, determinado plegamiento podría dificultar el inicio de la traducción, disminuyendo la cantidad de proteínas sintetizadas (Kubo e Imanaka 1989; Makino et al. 1997). Por lo tanto el UCS está involucrado en la estructura y estabilidad del ARNm, y de la proteína a sintetizar (Cortazzo et al. 2002; Duan et al. 2003; Khrustalev y Barkovsky 2012; Weygand-Durasevic e Ibba 2010). De esta forma, si las regiones TM "prefieren" a los codones que no poseen a los ARNt isoaceptores más abundantes de la célula, estos codones terminados en G pueden estar del alguna forma modulando la cinética de la traducción de esta región particular de las proteínas de membrana. Como se ha mencionado anteriormente la opción de codones que son reconocidos por ARNt poco abundantes genera una

disminución en la traducción (Sørensen et al. 1989). Sin embargo nuestro conocimiento sobre el patrón en el UCS en las proteínas de membrana es aún escaso.

Las regiones TM presentan un mayor contenido de GC en la tercera posición del codón respecto a las regiones NTM (Figura 21). La tercera base de los codones es la que posee más libertad de cambio. Este es un aspecto que podría ser importante ya que el aumento de GC en las regiones de lectura abierta en el ARNm conlleva a la estabilización de estructuras secundarias, afectando la eficacia de la traducción (Kubo e Imanaka 1989), el nivel de expresión de un gen, la estabilidad del ARNm y en la estructura de iniciación del ARNm como en el plegamiento de la proteína (Fredrick e Ibba 2010; Marin 2008).

El enriquecimiento en U en los codones de los aminoácidos que componen las regiones TM, podría estar vinculado no sólo con la hidrofobicidad de los aminoácidos sino también con la estructura secundaria de la proteína que codifique (Chiusano et al. 2000). Esta idea ha sido llevada al extremo por Prilusky y Bibi (2009) que han llegado a proponer que se puede predecir a las proteínas de membrana en función del contenido de U en el ARNm. Por otra parte, se especula que los ARNm de las proteínas de membrana en su origen fueron ricos en U, aumentando en el transcurso del tiempo la frecuencia de G, lo que le aportaría una mayor estabilidad al ARNm (Prilusky y Bibi 2009). Dadas las limitaciones impuestas por el código genético, el lugar donde los ARNm de las proteínas de membrana pueden enriquecerse en G es en la tercera posición.

El uso más frecuente de los codones terminados en G en los aminoácidos hidrofóbicos y ciertos aminoácidos hidrofílicos en las regiones TM, podría estar además asociado a la eficiencia de traducción y/o con algún tipo de reconocimiento estructural que lleve a su vez a una localización específica a nivel intracelular. Tampoco se puede descartar que la formación de determinadas estructuras secundarias a nivel del ARNm sea otro factor que regule la cantidad de proteína a ser sintetizada por lo secuencia mensajera en particular (Shabalina et al. 2013).

Las regiones TM tienden a presentar un menor contenido de GC total con respecto a las regiones NTM (Figura 20). Si bien las diferencias son significativas, son solamente de 2%. Esto es esperable ya que las regiones TM poseen un enriquecimiento en aminoácidos hidrofóbicos y estos poseen en las segunda base U. Por otra parte, las regiones NTM tendrían un mayor contenido en GC en la primera y segunda posición de los codones con respecto a las mismas posiciones en las regiones TM (Figura 21, Tabla A5). El contenido de GC3 es superior en las regiones TM con respecto a las regiones NTM (W, $p<0,05$). Sin embargo cuando se realiza un análisis más detallado en la composición de bases de las dos regiones, se encuentra que C3 no presenta diferencias significativas entre ambas, mientras que G3 está presente en mayor cantidad en las regiones TM (W, $p<0,05$) (Figura 22, Tabla A5).

El aumento del contenido de G3 en las regiones TM podría estar implicado en una compensación entre las regiones TM con respeto a las regiones NTM en la cantidad de GC; ya que si se suman el contenido de GC1 y GC2 las regiones NTM superan en un 10% aproximadamente, al contenido de GC1 y GC2 en las regiones TM. Se propone que el aumento de GC en la tercera posición en las regiones TM pueda estar implicado, como se mencionó anteriormente, en una forma de "compensación" a nivel de frecuencia de bases en estos genes.

A partir de los resultados obtenidos se puede decir que las regiones TM de las proteínas de membrana presentan un sesgo particular de codones sinónimos siendo preferidos los terminados en G. Ese patrón es fuerte y se encuentra en forma general conservado en la Familia. No es claro qué ventaja aporta dicha preferencia en el uso de codones sinónimos. Se ha asociado la fidelidad en la traducción de las proteínas de membrana y el correcto plegamiento de las regiones TM con regiones del ARNm en el cual presentan secuencias similares a la secuencia Shine–Dalgarno (Fluman et al. 2014). De esta manera se enlentece la traducción modelando la cinética del plegamiento de las regiones TM. Queda mucho para hacer todavía para poder establecer las fuerzas que están actuando y su magnitud en la participación del patrón observado.

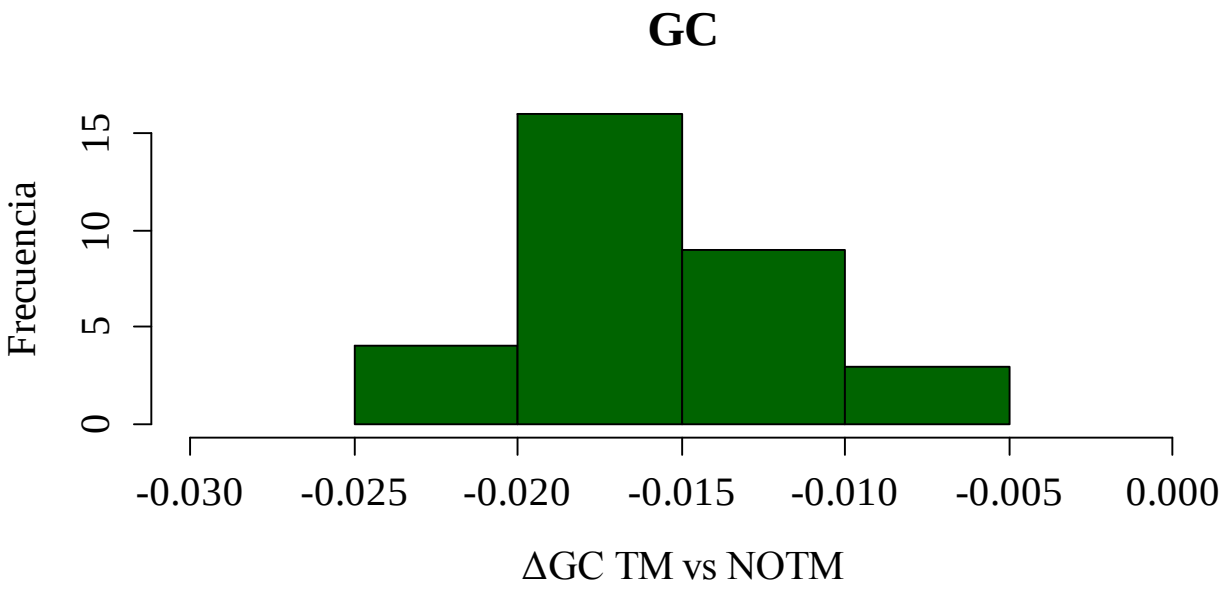

Figura 20. El ΔGC entre las regiones TM con respecto de las regiones NTM en los 32 organismos en estudio

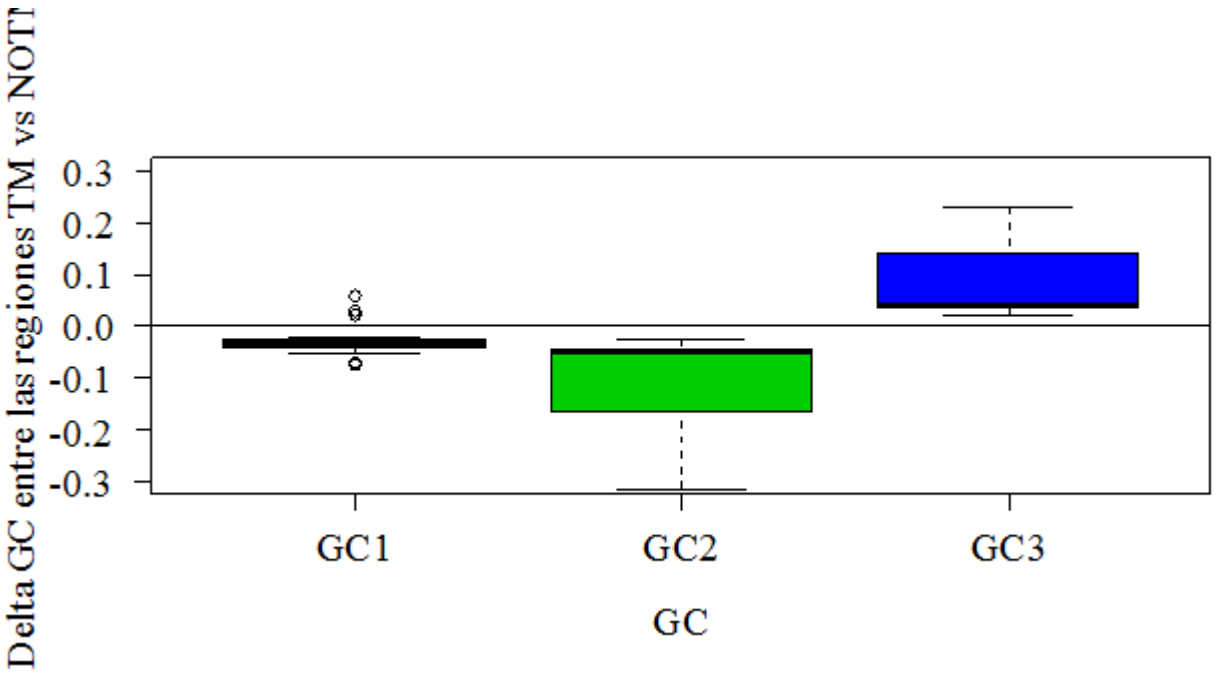

Figura 21. Diferencias en el contenido de GC en la primera, segunda y tercera posición de los codones de las regiones TM con respecto de las regiones NTM en los 32 organismos en estudio.

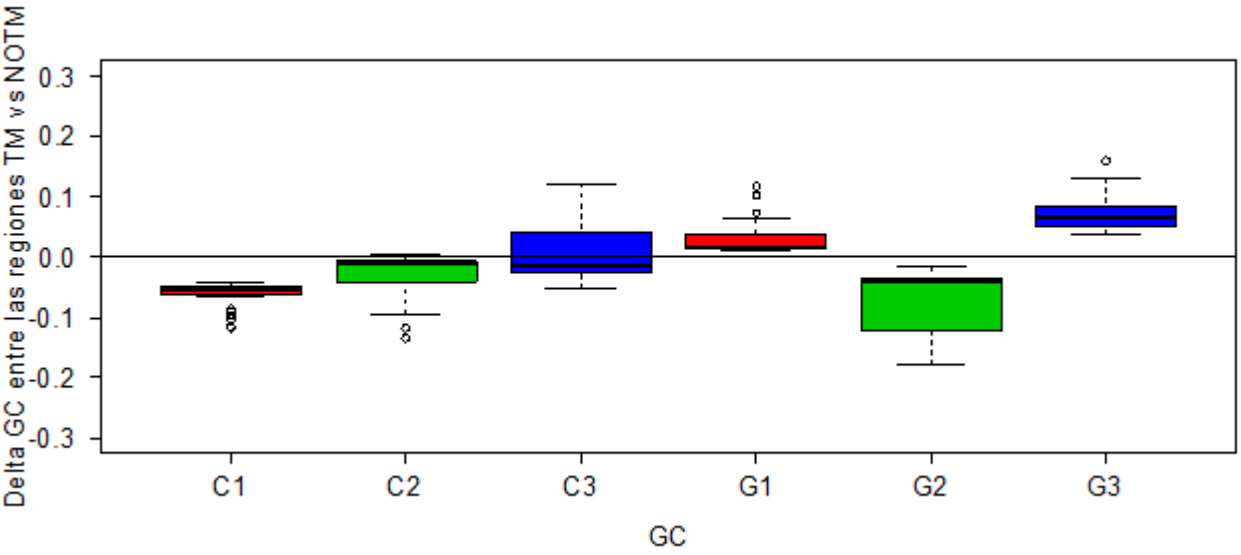

Figura 22. Diferencias en el contenido de G y C en la primera, segunda y tercera posición de los codones de las regiones TM con respecto de las regiones NTM en los 32 organismos en estudio.

aa	codon	Comamonadaceae bacterium CR	Rhodoferax ferrireducens T118	Hylemonella gracilis ATCC 19624	Verminephrobacter eiseniae EF01 2	Verminephrobacter At4	Acidovorax CF316	Acidovorax NO 1	Acidovorax radicis N35	Acidovorax KKS102	Acidovorax delafieldii 2AN	Delftia acidovorans SPH 1	Delftia Cs1 4	Delftia sp. JD2	Comamonas testosteroni S44	Comamonas testosteroni CNB 2	Comamonas testosteroni ATCC 11996	Comamonas testosteroni KF 1	Acidovorax JS42	Acidovorax ebreus TPSY	Alicycliphilus denitrificans BC	Alicycliphilus denitrificans K601	Acidovorax avenae ATCC 19860	Acidovorax citrulli AAC00 1	Variovorax paradoxus B4	Variovorax paradoxus S110	Variovorax paradoxus EPS	Variovorax CF313	Ramlibacter tataouinensis TTB310	Hydrogenophaga PBC	Polaromonas naphthalenivorans	Polaromonas JS666	Polaromonas CF318
F	UUU	x	+	-	x	+	x	x	-	-	x	-	-	-	x	x	x	-	+	x	x	-	+	-	-	-	+	x	x	x	-	x	+
F	UUC	x	+	-	x	+	x	x	-	-	x	-	-	-	x	x	x	-	+	x	x	-	+	-	-	-	+	x	x	x	-	x	+
Y	UAU	x	x	-	x	x	-	+	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	x	x	x	x	-	-	x	-	x
Y	UAC	x	x	-	x	x	-	+	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	x	x	x	x	-	-	x	-	x
K	AAA	-	+	-	x	-	x	x	x	+	x	+	+	+	+	-	-	x	x	-	x	x	+	-	-	-	-	x	+	x	-	x	+
K	AAG	-	+	-	x	-	x	x	x	+	x	+	+	+	+	-	-	x	x	-	x	x	+	-	-	-	-	x	+	x	-	x	+
C	UGU	x	-	-	x	-	x	+	x	x	x	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-	-	x	x	x	-	x	+
C	UGC	x	-	-	x	-	x	+	x	x	x	-	-	-	-	-	-	x	-	-	-	-	-	-	-	-	-	x	x	x	-	x	+
H	CAU	-	-	-	x	x	x	x	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	x	-	x	-
H	CAC	-	-	-	x	x	x	x	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	x	-	x	-
Q	CAA	-	-	-	x	-	x	x	x	-	x	+	+	+	-	x	-	x	-	-	-	-	-	-	-	-	x	+	x	x	x	-	x	+
Q	CAG	-	-	-	x	-	x	x	x	-	x	+	+	+	-	x	-	x	-	-	-	-	-	-	-	-	x	+	x	x	x	-	x	+
N	AAU	x	x	-	x	x	x	x	-	-	-	-	-	-	x	x	x	x	-	-	x	-	-	-	-	-	-	-	x	x	x	-	x	-
N	AAC	x	x	-	x	x	x	x	-	-	-	-	-	-	x	x	x	x	-	-	x	-	-	-	-	-	-	-	x	x	x	-	x	-
D	GAU	-	-	-	x	-	x	x	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	x	-	x	-
D	GAC	-	-	-	x	-	x	x	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	x	-	x	-
E	GAA	-	-	-	x	-	x	x	-	+	-	x	x	x	-	-	-	x	-	-	-	-	-	-	-	+	x	x	x	-	x	-	x	-
E	GAG	-	-	-	x	-	x	x	-	+	-	x	x	x	-	-	-	x	-	-	-	-	-	-	-	+	x	x	x	-	x	-	x	-
I	AUU	-	x	-	x	x	x	x	-	+	x	-	-	-	x	x	-	-	-	-	-	-	+	-	+	-	x	x	x	x	x	x	x
I	AUC	x	x	-	x	-	x	x	-	-	x	x	x	x	-	+	-	x	-	x	x	-	x	-	-	+	x	x	x	x	x	-	x	-
I	AUA	+	x	-	x	x	x	x	x	-	x	x	x	x	x	x	x	x	-	x	x	x	x	-	-	-	x	-	x	x	x	x	x	+
V	GUU	-	+	-	x	-	x	x	-	-	x	x	x	x	-	x	-	x	+	-	x	x	-	+	-	+	-	x	x	x	x	x	x
V	GUC	x	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	-	+	-	-	x	x	x	x	-	x	x	-	x
V	GUA	+	+	x	x	+	x	x	x	x	x	x	x	x	x	x	x	x	x	+	+	-	x	+	-	-	-	x	x	x	x	x	-
V	GUG	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x
P	CCU	-	-	x	x	-	x	x	+	x	x	x	x	x	-	x	x	x	x	x	x	+	x	x	-	x	-	x	x	x	+	x	x
P	CCC	x	x	-	x	x	x	x	x	x	x	-	+	-	-	x	-	x	x	x	x	x	+	-	x	x	x	x	x	x	x	x	x
P	CCA	+	+	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	+	x	x	x	x	x	x	x	x	x	x	x	x	x
P	CCG	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
T	ACU	+	-	-	x	-	x	x	-	-	x	x	x	x	+	x	x	x	x	x	x	+	x	-	-	-	-	-	x	x	-	x	x
T	ACC	x	-	-	x	x	x	x	x	x	x	-	+	-	-	x	-	x	x	x	-	-	+	+	x	x	x	x	x	x	+	x	x
T	ACA	-	x	+	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
T	ACG	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
A	GCU	x	-	x	x	-	x	x	-	-	x	x	x	x	-	x	+	x	+	-	x	-	x	x	-	-	-	x	x	x	-	x	-
A	GCC	x	+	-	x	x	x	x	x	x	x	x	x	x	-	x	-	x	x	x	-	-	+	x	x	x	x	x	x	x	x	x	x
A	GCA	x	x	+	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
A	GCG	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
G	GGU	-	x	-	x	x	x	x	x	x	x	x	+	x	x	-	x	+	-	-	-	-	-	-	+	x	-	x	-	-	x	x	x	+
G	GGC	x	-	-	x	-	x	x	x	x	x	x	-	x	-	x	-	x	-	-	-	-	-	-	-	x	x	x	x	x	x	-	x	+
G	GGA	-	x	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	x	x	x	x
G	GGG	x	-	-	x	-	x	+	x	x	x	x	x	x	-	x	-	x	x	-	+	x	x	x	x	x	x	x	x	x	-	x	x
L	UUA	-	-	-	x	-	x	x	-	-	x	-	+	-	-	x	x	x	x	-	-	-	x	-	-	-	-	x	x	x	x	x	-
L	UUG	-	-	x	x	x	x	x	-	-	x	x	+	x	-	x	-	-	x	x	x	x	x	x	x	-	-	x	x	-	-	x	+
L	CUU	+	x	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	-	-	-	x	x	x	x	x	x
L	CUC	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	+	x	x	x	x	x	x	x	x	x
L	CUA	-	-	-	x	-	x	x	x	x	x	x	x	x	x	x	x	x	-	-	-	-	x	x	-	+	-	x	x	x	x	x	+
L	CUG	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
R	CGU	+	-	x	-	-	x	+	+	-	-	+	-	-	-	-	-	-	x	-	-	-	x	+	-	-	-	x	x	x	-	x	-
R	CGC	+	-	+	x	-	x	x	-	x	x	+	-	+	x	x	-	x	-	-	-	+	-	-	-	-	-	x	x	x	x	x	-
R	CGA	-	-	-	x	-	x	x	-	-	x	-	+	-	-	x	-	x	-	-	-	+	-	+	-	-	-	x	x	x	-	x	-
R	CGG	-	-	-	-	-	x	-	-	x	+	-	-	-	-	-	-	x	-	-	-	-	-	-	+	+	+	x	-	x	x	-	x	-
R	AGA	-	-	-	x	-	x	x	-	-	x	-	-	-	-	x	-	x	-	-	+	-	-	-	+	+	-	x	x	x	x	-	x	-
R	AGG	-	-	-	x	-	x	x	-	-	x	x	x	x	x	x	x	x	+	-	-	-	-	+	+	-	-	x	x	x	x	x	x
S	UCU	-	-	-	x	-	x	x	+	+	x	x	+	x	+	x	-	x	x	+	x	x	x	x	-	-	-	x	x	x	-	x	+
S	UCC	+	-	-	x	-	x	x	x	x	-	-	-	-	-	x	x	x	x	x	-	-	x	x	x	x	-	x	x	x	-	x	x
S	UCA	-	-	x	x	-	x	x	-	+	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	+	x	x	x	x	x	x
S	UCG	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
S	AGU	-	-	+	x	+	x	x	+	-	x	+	x	x	-	x	+	x	-	-	x	-	x	x	+	-	-	x	x	x	+	x	-
S	AGC	x	x	x	x	x	x	-	x	x	-	x	x	x	x	x	x	+	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Figura 23. Codones preferidos y evitados por las regiones TM con respecto a las regiones NTM, en la familia *Comamonadaceae*. Se utilizó el estadístico $\chi 2$ de contingencia para poder determinar que codón es preferido o evitado en forma significativa por las regiones ($p < 0,05$). En rojo y bordó estan representados los codones preferidos por las regiones TM ($p<0,01$) y ($p<0,05$) respectivamente, mientras que en azul y violeta estan los codones evitados o (Preferidos por las regiones NTM) ($p<0,01$ y $p< 0,05$) respectivamente.

CONCLUSIÓN

Un sesgo selectivo a nivel de expresión en el uso de codones en un genoma se puede definir como significativo si cumple con tres criterios: 1) la tendencia principal en la variación en el uso de codones debe ser explicado por la expresión, en este caso, esto se muestra mediante una correlación significativa entre el primer o el segundo eje generado por el análisis WCA y el MELP, usado este último como predictor de nivel de expresión; 2) tiene un coeficiente significativo selectivo (S), y 3) genes altamente expresados deben de caracterizarse por una tasa de sustitución sinónima baja, esto se muestra por una correlación negativa significativa entre la estimación del dS y el valor MELP.

17 organismos cumplieron con estas tres condiciones simultáneamente, las cuatro cepas de *Comamonas testosteroni*, *Ramlibacter tataouinensis* TTB310, las tres cepas de *Delftia*, las dos cepas de *Variovorax paradoxus*: B4 y S110, *Verminephrobacter* sp. At4, cinco cepas de de *Acidovorax*: *A.* sp. NO 1, *A.* sp. KKS102, *A. avenae* ATCC 19860, *A. citrulli* AAC00 1, *A. delafieldii* 2, *Polaromonas* CF318.

Entre estos organismos, las cepas pertenecientes a las *C. testosteroni* y las *Delftia* son las que presentan la tendencia más clara y fuerte, y por esta razón se cree que estas bacterias presentan el sesgo selectivo en el uso de codones más fuerte. Dado que estas bacterias forman un grupo monofilético, el incremento debe ser entendido como un cambio reciente en el linaje.

Otras bacterias que no cumplieron estrictamente con los tres criterios definidos previamente, también presentan tendencias en el uso de codones que podrían ser fácilmente explicadas en términos de selección para traducción, este es el caso de *A. radicis* N35, *A.* sp. CF316, *P.* sp. JS666, *V. eiseniae* EF01 2, *V. paradoxus* EPS y *H. gracilis* ATCC 19624.

Ningún resultado apoyó el uso selectivo en el sesgo de los codones para *Rhodoferax ferrireducens* T118, *Comamonadaceae bacterium* CR, las dos cepas de *Alicycliphilus* e *Hydrogenophaga* sp. PBC. En estos casos, la mutación y la deriva genética probablemente hayan borrado el efecto de la selección que actúa sobre los genes altamente expresados.

Nuestros resultados sugieren que la selección natural también es operativa en los niveles de precisión de la traducción en muchos organismos. Investigamos el sesgo en el uso de codones sinónimos en las regiones conservadas de las proteínas, que tiende a ser similar al sesgo observado en los GAE para la mayoría de los codones. También mostramos que existe una correlación significativa entre la divergencia de los cambios sinónimos y no sinónimos, apoyando la idea de que la selección en la precisión de la traducción es operativa. El sesgo selectivo es muy variable en la Familia, y es observado claramente para muchos de los organismos de este grupo. Como regla general, los organismos más estrechamente relacionados tienden a mostrar sesgos más similares. Sin embargo, codones óptimos similares pueden ser observados en organismos relativamente distantes, esto apoya la idea de que ha habido un grupo estable de codones preferidos, a pesar de que muchos linajes presentan un efecto relajado de la selección. La correlación entre la divergencia sinónima con el nivel de expresión (MELP), observada cuando se comparan organismos lejanamente relacionados, también apoya esta conclusión. Algunos casos interesantes, con sesgos divergentes también pudieron ser identificados. La coincidencia observada entre los codones óptimos y el grupo de anticodones identificado en cada genoma también puede explicarse en términos de selección. La conservación de los grupos de los anticodones también se observa en diferentes linajes, lo que de alguna manera explica el mantenimiento de los codones óptimos observados en diferentes linajes. Las diferencias observadas en el coeficiente de selección estimado podría reflejar en parte la heterogeneidad en los estilos de vida, tamaños efectivos poblacionales, tiempos de generación, nichos ecológicos, rutas metabólicas, etc., que caracterizan a los diferentes microorganismos de esta familia.

Los genes de las proteínas ribosomales presentan un contenido de GC y de energía mínima de plegamiento normalizada inferior con respecto al resto de los genes en todos los organismos. Sin embargo, esta energía mínima se desvía de lo esperado con respecto al contenido de GC. A partir de este resultado se puede inferir que el plegamiento en el ARNm de los genes de las proteínas ribosomales presenta una menor probabilidad de generar potenciales estructuras secundarias en comparación con el resto de los genes del genoma, facilitando la rápida asociación de la maquinaria de traducción.

Las regiones TM presentan un enriquecimiento en codones terminados en la base G. Estos codones son los tripletes menos frecuentes en los genes de alta expresión. De esta manera el sesgo en el uso de codones por las regiones TM podría estar implicado en el plegamiento del ARNm o algún tipo estrategia en la síntesis de estas proteínas y en particular de estas importantes regiones.

BIBLIOGRAFIA

Akashi, H. 1994. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics **136**(3): 927-935.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. Journal of molecular biology **215**(3): 403-410.

Amann, R., Ludwig, W., Schulze, R., Spring, S., Moore, E., and Schleifer, K.-H. 1996. rRNA-targeted oligonucleotide probes for the identification of genuine and former pseudomonads. Systematic and applied microbiology **19**(4): 501-509.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., and Kubal, M. 2008. The RAST Server: rapid annotations using subsystems technology. BMC genomics **9**(1): 1.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., and Prjibelski, A.D. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology **19**(5): 455-477.

Bar-Ness, E., Hadar, Y., Chen, Y., Römheld, V., and Marschner, H. 1992. Short-term effects of rhizosphere microorganisms on Fe uptake from microbial siderophores by maize and oat. Plant Physiology **100**(1): 451-456.

Boyd, D., Schierle, C., and Beckwith, J. 1998. How many membrane proteins are there? Protein science : a publication of the Protein Society **7**(1): 201-205. doi: 10.1002/pro.5560070121.

Bulmer, M. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics **129**(3): 897-907.

Cannarozzi, G., Schraudolph, N.N., Faty, M., von Rohr, P., Friberg, M.T., Roth, A.C., Gonnet, P., Gonnet, G., and Barral, Y. 2010. A role for codon order in translation dynamics. Cell **141**(2): 355-367. doi: 10.1016/j.cell.2010.02.036.

Castro-Sowinski, S., Herschkovitz, Y., Okon, Y., and Jurkevitch, E. 2007. Effects of inoculation with plant growth-promoting rhizobacteria on resident rhizosphere microorganisms. FEMS microbiology letters **276**(1): 1-11.

Claflin, L., Ramundo, B., Leach, J., and Erinle, I. 1989. Pseudomonas avenae, causal agent of bacterial leaf stripe of pearl millet. Plant disease **73**(12): 1010-1014.

Cordero, O.X., and Hogeweg, P. 2009. The impact of long-distance horizontal gene transfer on prokaryotic genome size. Proceedings of the National Academy of Sciences **106**(51): 21748-21753.

Cortazzo, P., Cervenansky, C., Marin, M., Reiss, C., Ehrlich, R., and Deana, A. 2002. Silent mutations affect in vivo protein folding in Escherichia coli. Biochemical and biophysical research communications **293**(1): 537-541. doi: 10.1016/S0006-291X(02)00226-7.

Curran, A.R., and Engelman, D.M. 2003. Sequence motifs, polar interactions and conformational changes in helical membrane proteins. Current opinion in structural biology **13**(4): 412-417.

Charif, D., Thioulouse, J., Lobry, J., and Perrière, G. 2005. Online synonymous codon usage analyses with the ade4 and seqinR packages. Bioinformatics **21**(4): 545-547.

Chen, S.L., Lee, W., Hottes, A.K., Shapiro, L., and McAdams, H.H. 2004. Codon usage between genomes is constrained by genome-wide mutational processes. Proceedings of the National Academy of Sciences of the United States of America **101**(10): 3480-3485. doi: 10.1073/pnas.0307827100.

Chen, W.-M., Lin, Y.-S., Sheu, D.-S., and Sheu, S.-Y. 2012. Delftia litopenaei sp. nov., a poly-β-hydroxybutyrate-accumulating bacterium isolated from a freshwater shrimp culture pond. International journal of systematic and evolutionary microbiology **62**(10): 2315-2321.

Cheung, K., and Gu, J.-D. 2007. Mechanism of hexavalent chromium detoxification by microorganisms and bioremediation application potential: a review. International Biodeterioration & Biodegradation **59**(1): 8-15.

Chiusano, M.L., Alvarez-Valin, F., Di Giulio, M., D'Onofrio, G., Ammirato, G., Colonna, G., and Bernardi, G. 2000. Second codon positions of genes and the secondary structures of proteins.

Relationships and implications for the origin of the genetic code. Gene **261**(1): 63-69.
Chursov, A., Frishman, D., and Shneider, A. 2013. Conservation of mRNA secondary structures may filter out mutations in Escherichia coli evolution. Nucleic acids research: gkt507.
Dawson, J.P., Weinger, J.S., and Engelman, D.M. 2002. Motifs of serine and threonine can drive association of transmembrane helices. Journal of molecular biology **316**(3): 799-805. doi: 10.1006/jmbi.2001.5353.
De Vos, P., Kersters, K., Falsen, E., Pot, B., Gillis, M., Segers, P., and De Ley, J. 1985. Comamonas Davis and Park 1962 gen. nov., nom. rev. emend., and Comamonas terrigena Hugh 1962 sp. nov., nom. rev. International Journal of Systematic and Evolutionary Microbiology **35**(4): 443-453.
Del Campo, C., Bartholomäus, A., Fedyunin, I., and Ignatova, Z. 2015. Secondary structure across the bacterial transcriptome reveals versatile roles in mRNA regulation and function. PLoS genetics **11**(10): e1005613.
Duan, J., Wainwright, M.S., Comeron, J.M., Saitou, N., Sanders, A.R., Gelernter, J., and Gejman, P.V. 2003. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Human molecular genetics **12**(3): 205-216.
Faure, G., Ogurtsov, A.Y., Shabalina, S.A., and Koonin, E.V. 2016. Role of mRNA structure in the control of protein folding. Nucleic acids research: gkw671.
Felsenstein, J. 1985. Phylogenies and the comparative method. American Naturalist: 1-15.
Fluman, N., Navon, S., Bibi, E., and Pilpel, Y. 2014. mRNA-programmed translation pauses in the targeting of E. coli membrane proteins. Elife **3**: e03440.
Francino, M.P., and Ochman, H. 2001. Deamination as the basis of strand-asymmetric evolution in transcribed Escherichia coli sequences. Molecular biology and evolution **18**(6): 1147-1150.
Frank, A., and Lobry, J. 1999. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene **238**(1): 65-77.
Frederico, L.A., Kunkel, T.A., and Shaw, B.R. 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry **29**(10): 2532-2537.
Fredrick, K., and Ibba, M. 2010. How the sequence of a gene can tune its translation. Cell **141**(2): 227-229. doi: 10.1016/j.cell.2010.03.033.
Garavaglia, L., Cerdeira, S.B., and Vullo, D.L. 2010. Chromium (VI) biotransformation by beta- and gamma-Proteobacteria from natural polluted environments: a combined biological and chemical treatment for industrial wastes. Journal of hazardous materials **175**(1-3): 104-110. doi: 10.1016/j.jhazmat.2009.09.134.
Gaucher, E.A., De Kee, D.W., and Benner, S.A. 2006. Application of DETECTER, an evolutionary genomic tool to analyze genetic variation, to the cystic fibrosis gene family. BMC genomics **7**: 44. doi: 10.1186/1471-2164-7-44.
Goetz, R.M., and Fuglsang, A. 2005. Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from Escherichia coli. Biochemical and biophysical research communications **327**(1): 4-7.
Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., and Tiedje, J.M. 2007. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. International journal of systematic and evolutionary microbiology **57**(1): 81-91.
Gorochowski, T.E., Ignatova, Z., Bovenberg, R.A., and Roubos, J.A. 2015. Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate. Nucleic acids research: gkv199.
Gouy, M., and Gautier, C. 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic acids research **10**(22): 7055-7074.
Goyal, A.K., and Zylstra, G.J. 1996. Molecular cloning of novel genes for polycyclic aromatic hydrocarbon degradation from Comamonas testosteroni GZ39. Applied and environmental microbiology **62**(1): 230-236.
Grantham, R., Gautier, C., and Gouy, M. 1980. Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. Nucleic acids research **8**(9):

1893-1912.
Grosjean, H., and Fiers, W. 1982. Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene **18**(3): 199-209.
Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic biology **59**(3): 307-321.
Gustilo, E.M., Vendeix, F.A., and Agris, P.F. 2008. tRNA's modifications bring order to gene expression. Current opinion in microbiology **11**(2): 134-140.
Hershberg, R., and Petrov, D.A. 2008. Selection on codon bias. Annual review of genetics **42**: 287-299.
Hershberg, R., and Petrov, D.A. 2009. General rules for optimal codon choice. PLoS genetics **5**(7): e1000556.
Hessa, T., Kim, H., Bihlmaier, K., Lundin, C., Boekel, J., Andersson, H., Nilsson, I., White, S.H., and von Heijne, G. 2005. Recognition of transmembrane helices by the endoplasmic reticulum translocon. Nature **433**(7024): 377-381. doi: 10.1038/nature03216.
Hiraishi, A. 1994. Phylogenetic affiliations ofRhodoferax fermentans and related species of phototrophic bacteria as determined by automated 16S rDNA sequencing. Current microbiology **28**(1): 25-29.
Housby, J.N., and Southern, E.M. 1998. Fidelity of DNA ligation: a novel experimental approach based on the polymerisation of libraries of oligonucleotides. Nucleic acids research **26**(18): 4259-4266.
Ikemura, T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. Journal of molecular biology **151**(3): 389-409.
Ikemura, T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Molecular biology and evolution **2**(1): 13-34.
Ingolia, N.T. 2016. Ribosome Footprint Profiling of Translation throughout the Genome. Cell **165**(1): 22-33.
Ingolia, N.T., Ghaemmaghami, S., Newman, J.R., and Weissman, J.S. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science **324**(5924): 218-223.
Iriarte, A., Baraibar, J.D., Romero, H., Castro-Sowinski, S., and Musto, H. 2013. Evolution of optimal codon choices in the family Enterobacteriaceae. Microbiology **159**(3): 555-564.
Iriarte, A., Baraibar, J.D., Romero, H., and Musto, H. 2011. Selected codon usage bias in members of the class Mollicutes. Gene **473**(2): 110-118. doi: 10.1016/j.gene.2010.11.010.
Iriarte, A., Jara, E., Leytón, L., Diana, L., and Musto, H. 2014. General Trends in Selectively Driven Codon Usage Biases in the Domain Archaea. Journal of molecular evolution **79**(3-4): 105-110.
Juarez-Jimenez, B., Manzanera, M., Rodelas, B., Martinez-Toledo, M.V., Gonzalez-Lopez, J., Crognale, S., Pesciaroli, C., and Fenice, M. 2010. Metabolic characterization of a strain (BM90) of Delftia tsuruhatensis showing highly diversified capacity to degrade low molecular weight phenols. Biodegradation **21**(3): 475-489. doi: 10.1007/s10532-009-9317-4.
Kahali, B., Basak, S., and Ghosh, T.C. 2008. Delving deeper into the unexpected correlation between gene expressivity and codon usage bias of Escherichia coli genome. Journal of Biomolecular Structure and Dynamics **25**(6): 655-661.
Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y., and Ikemura, T. 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. Journal of molecular evolution **53**(4-5): 290-298.
Kanaya, S., Yamada, Y., Kudo, Y., and Ikemura, T. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene **238**(1): 143-155.

Kanazawa, S., and Mori, K. 1996. Isolation of Cadmium-Resistant Bacteria and Their Resistance Mechanisms: Part 1. Isolation of Cd-Resistant Bacteria from Soils Contaminated with Heavy Metals. Soil science and plant nutrition **42**(4): 725-730.
Keane, T.M., Creevey, C.J., Pentony, M.M., Naughton, T.J., and Mclnerney, J.O. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC evolutionary biology **6**(1): 1.
Khrustalev, V.V., and Barkovsky, E.V. 2012. Stabilization of secondary structure elements by specific combinations of hydrophilic and hydrophobic amino acid residues is more important for proteins encoded by GC-poor genes. Biochimie **94**(12): 2706-2715.
Kim, M., Oh, H.-S., Park, S.-C., and Chun, J. 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. International journal of systematic and evolutionary microbiology **64**(2): 346-351.
Kimchi-Sarfaty, C., Oh, J.M., Kim, I.W., Sauna, Z.E., Calcagno, A.M., Ambudkar, S.V., and Gottesman, M.M. 2007. A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science **315**(5811): 525-528. doi: 10.1126/science.1135308.
Kimura, M. 1968. Evolutionary rate at the molecular level. Nature **217**(5129): 624-626.
King, J.L., and Jukes, T.H. 1969. Non-Darwinian evolution. Science **164**(3881): 788-798.
Kjeldsen, K.U., Bataillon, T., Pinel, N., De Mita, S., Lund, M.B., Panitz, F., Bendixen, C., Stahl, D.A., and Schramm, A. 2012. Purifying selection and molecular adaptation in the genome of Verminephrobacter, the heritable symbiotic bacteria of earthworms. Genome biology and evolution **4**(3): 307-315.
Knight, R.D., Freeland, S.J., and Landweber, L.F. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome biology **2**(4): 1.
Koonin, E.V., and Wolf, Y.I. 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic acids research **36**(21): 6688-6719. doi: 10.1093/nar/gkn668.
Krogh, A., Larsson, B., Von Heijne, G., and Sonnhammer, E.L. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of molecular biology **305**(3): 567-580.
Kubo, M., and Imanaka, T. 1989. mRNA secondary structure in an open reading frame reduces translation efficiency in Bacillus subtilis. Journal of bacteriology **171**(7): 4080-4082.
Kudla, G., Murray, A.W., Tollervey, D., and Plotkin, J.B. 2009. Coding-sequence determinants of gene expression in Escherichia coli. Science **324**(5924): 255-258. doi: 10.1126/science.1170160.
Lawrence, J. 1999. Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. Current opinion in genetics & development **9**(6): 642-648.
Lawrence, J.G. 1997. Selfish operons and speciation by gene transfer. Trends in microbiology **5**(9): 355-359.
Leadbetter, J.R., and Greenberg, E. 2000. Metabolism of acyl-homoserine lactone quorum-sensing signals by Variovorax paradoxus. Journal of Bacteriology **182**(24): 6921-6926.
Liu, Y., Engelman, D.M., and Gerstein, M. 2002. Genomic analysis of membrane protein families: abundance and conserved motifs. Genome biology **3**(10): research0054.
Lobry, J.R. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Molecular biology and evolution **13**(5): 660-665.
Lorenz, R., Bernhart, S.H., Zu Siederdissen, C.H., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. 2011. ViennaRNA Package 2.0. Algorithms for Molecular Biology **6**(1): 1.
Lott, B.B., Wang, Y., and Nakazato, T. 2013. A comparative study of ribosomal proteins: linkage between amino acid distribution and ribosomal assembly. BMC biophysics **6**(1): 1.
Lowe, T.M., and Eddy, S.R. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic acids research **25**(5): 955-964.
Lund, M.B., Davidson, S.K., Holmstrup, M., James, S., Kjeldsen, K.U., Stahl, D.A., and Schramm, A. 2010. Diversity and host specificity of the Verminephrobacter-earthworm symbiosis. Environmental microbiology **12**(8): 2142-2151. doi: 10.1111/j.1462-2920.2009.02084.x.

Makino, S., Qu, J.N., Uemori, K., Ichikawa, H., Ogura, T., and Matsuzawa, H. 1997. A silent mutation in the ftsH gene of Escherichia coli that affects FtsH protein production and colicin tolerance. Molecular & general genetics : MGG **254**(5): 578-583.
Marin, M. 2008. Folding at the rhythm of the rare codon beat. Biotechnology journal **3**(8): 1047-1057. doi: 10.1002/biot.200800089.
Martinez-Gil, L., Sauri, A., Marti-Renom, M.A., and Mingarro, I. 2011. Membrane protein integration into the endoplasmic reticulum. The FEBS journal **278**(20): 3846-3858. doi: 10.1111/j.1742-4658.2011.08185.x.
Melnyk, R.A., Kim, S., Curran, A.R., Engelman, D.M., Bowie, J.U., and Deber, C.M. 2004. The affinity of GXXXG motifs in transmembrane helix-helix interactions is modulated by long-range communication. The Journal of biological chemistry **279**(16): 16591-16597. doi: 10.1074/jbc.M313936200.
Mergaert, J., and Swings, J. 1996. Biodiversity of microorganisms that degrade bacterial and synthetic polyesters. Journal of industrial microbiology **17**(5-6): 463-469.
Mira, A., Ochman, H., and Moran, N.A. 2001. Deletional bias and the evolution of bacterial genomes. Trends in Genetics **17**(10): 589-596.
Mishra, D., and Kar, M. 1974. Nickel in plant growth and metabolism. The botanical review **40**(4): 395-452.
Morel, M.A., Ubalde, M.C., Brana, V., and Castro-Sowinski, S. 2011. Delftia sp. JD2: a potential Cr(VI)-reducing agent with plant growth-promoting activity. Archives of microbiology **193**(1): 63-68. doi: 10.1007/s00203-010-0632-2.
Moura, A., Savageau, M.A., and Alves, R. 2013. Relative amino acid composition signatures of organisms and environments. PloS one **8**(10): e77319.
Musto, H., Naya, H., Zavala, A., Romero, H., Alvarez-Valin, F., and Bernardi, G. 2006. Genomic GC level, optimal growth temperature, and genome size in prokaryotes. Biochemical and biophysical research communications **347**(1): 1-3. doi: 10.1016/j.bbrc.2006.06.054.
Musto, H., Romero, H., and Zavala, A. 2003. Translational selection is operative for synonymous codon usage in Clostridium perfringens and Clostridium acetobutylicum. Microbiology **149**(Pt 4): 855-863. doi: 10.1099/mic.0.26063-0.
Muto, A., and Osawa, S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. Proceedings of the National Academy of Sciences of the United States of America **84**(1): 166-169.
Naya, H., Romero, H., Zavala, A., Alvarez, B., and Musto, H. 2002. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. Journal of molecular evolution **55**(3): 260-264. doi: 10.1007/s00239-002-2323-3.
Nuñez, P.A., Romero, H., Farber, M.D., and Rocha, E.P. 2013. Natural selection for operons depends on genome size. Genome biology and evolution **5**(11): 2242-2254.
Paradis, E., Claude, J., and Strimmer, K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics **20**(2): 289-290.
Park, C., Chen, X., Yang, J.-R., and Zhang, J. 2013. Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. Proceedings of the National Academy of Sciences **110**(8): E678-E686.
Pop, C., Rouskin, S., Ingolia, N.T., Han, L., Phizicky, E.M., Weissman, J.S., and Koller, D. 2014. Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. Molecular systems biology **10**(12): 770.
Precup, J., and Parker, J. 1987. Missense misreading of asparagine codons as a function of codon identity and context. The Journal of biological chemistry **262**(23): 11351-11355.
Prilusky, J., and Bibi, E. 2009. Studying membrane proteins through the eyes of the genetic code revealed a strong uracil bias in their coding mRNAs. Proceedings of the National Academy of Sciences of the United States of America **106**(16): 6662-6666. doi: 10.1073/pnas.0902029106.
Pruitt, K.D., Tatusova, T., and Maglott, D.R. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research

33(suppl 1): D501-D504.
Ran, W., Kristensen, D.M., and Koonin, E.V. 2014. Coupling between protein level selection and codon usage optimization in the evolution of bacteria and archaea. MBio **5**(2): e00956-00914.
Revell, L.J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution **3**(2): 217-223.
Rice, P., Longden, I., and Bleasby, A. 2000. EMBOSS: the European molecular biology open software suite. Trends in genetics **16**(6): 276-277.
Rocha, E.P. 2004a. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome research **14**(11): 2279-2286. doi: 10.1101/gr.2896904.
Rocha, E.P. 2004b. The replication-related organization of bacterial genomes. Microbiology **150**(Pt 6): 1609-1627. doi: 10.1099/mic.0.26974-0.
Rocha, E.P. 2008. Evolutionary patterns in prokaryotic genomes. Current opinion in microbiology **11**(5): 454-460.
Santos, M.A., Moura, G., Massey, S.E., and Tuite, M.F. 2004. Driving change: the evolution of alternative genetic codes. TRENDS in Genetics **20**(2): 95-102.
SCANDOLA*, M., Finelli, L., Sarti, B., Mergaert, J., Swings, J., Ruffieux, K., Wintermantel, E., Boelens, J., De Wilde, B., and Müller, W.-R. 1998. Biodegradation of a starch containing thermoplastic in standardized test systems. Journal of Macromolecular Science, Part A: Pure and Applied Chemistry **35**(4): 589-608.
Seffens, W., and Digby, D. 1999. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. Nucleic acids research **27**(7): 1578-1584.
Shabalina, S.A., Spiridonov, N.A., and Kashina, A. 2013. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic acids research **41**(4): 2073-2094. doi: 10.1093/nar/gks1205.
Sharp, P.M., Bailes, E., Grocock, R.J., Peden, J.F., and Sockett, R.E. 2005. Variation in the strength of selected codon usage bias among bacteria. Nucleic acids research **33**(4): 1141-1153. doi: 10.1093/nar/gki242.
Sharp, P.M., Emery, L.R., and Zeng, K. 2010. Forces that influence the evolution of codon bias. Philosophical Transactions of the Royal Society B: Biological Sciences **365**(1544): 1203-1212.
Sharp, P.M., and Li, W.-H. 1986. Codon usage in regulatory genes in Escherichia coli does not reflect selection for 'rare'codons. Nucleic acids research **14**(19): 7737-7749.
Sonnhammer, E.L., Von Heijne, G., and Krogh, A. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. *In* Ismb. pp. 175-182.
Sørensen, M.A., Kurland, C., and Pedersen, S. 1989. Codon usage determines translation rate in Escherichia coli. Journal of molecular biology **207**(2): 365-377.
Stoletzki, N. 2008. Conflicting selection pressures on synonymous codon use in yeast suggest selection on mRNA secondary structures. BMC Evolutionary Biology **8**(1): 1.
Stoletzki, N., and Eyre-Walker, A. 2007. Synonymous codon usage in Escherichia coli: selection for translational accuracy. Molecular biology and evolution **24**(2): 374-381.
Stoppel, R.d., and Schlegel, H. 1995. Nickel-resistant bacteria from anthropogenically nickel-polluted and naturally nickel-percolated ecosystems. Applied and environmental microbiology **61**(6): 2276-2285.
Supek, F., and Vlahoviček, K. 2004. INCA: synonymous codon usage analysis and clustering by means of self-organizing map. Bioinformatics **20**(14): 2329-2330.
Supek, F., and Vlahoviček, K. 2005. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. BMC bioinformatics **6**(1): 1.
Suzuki, H., Brown, C.J., Forney, L.J., and Top, E.M. 2008. Comparison of correspondence analysis methods for synonymous codon usage in bacteria. DNA research **15**(6): 357-365.
Taboada, B., Ciria, R., Martinez-Guerrero, C.E., and Merino, E. 2012. ProOpDB: prokaryotic operon database. Nucleic acids research **40**(D1): D627-D631.
Talavera, G., and Castresana, J. 2007. Improvement of phylogenies after removing divergent and

ambiguously aligned blocks from protein sequence alignments. Systematic biology **56**(4): 564-577.
Thanaraj, T., and Argos, P. 1996. Protein secondary structural types are differentially coded on messenger RNA. Protein science **5**(10): 1973-1983.
Thioulouse, J., Chessel, D., Dole, S., and Olivier, J.-M. 1997. ADE-4: a multivariate analysis and graphical display software. Statistics and computing **7**(1): 75-83.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research **22**(22): 4673-4680.
Treangen, T.J., and Rocha, E.P. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS genetics **7**(1): e1001284. doi: 10.1371/journal.pgen.1001284.
Trotta, E. 2014. On the normalization of the minimum free energy of RNAs by sequence length. PloS one **9**(11): e113380. doi: 10.1371/journal.pone.0113380.
Trylska, J., Konecny, R., Tama, F., Brooks, C.L., and McCammon, J.A. 2004. Ribosome motions modulate electrostatic properties. Biopolymers **74**(6): 423-431.
Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell **141**(2): 344-354. doi: 10.1016/j.cell.2010.03.031.
Varenne, S., Buc, J., Lloubes, R., and Lazdunski, C. 1984. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. Journal of molecular biology **180**(3): 549-576.
Vieira-Silva, S., and Rocha, E.P. 2008. An assessment of the impacts of molecular oxygen on the evolution of proteomes. Molecular biology and evolution **25**(9): 1931-1942.
Wallin, E., and von Heijne, G. 1998. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. Protein science : a publication of the Protein Society **7**(4): 1029-1038. doi: 10.1002/pro.5560070420.
Warner, J.R. 1999. The economics of ribosome biosynthesis in yeast. Trends in biochemical sciences **24**(11): 437-440.
Wen, A., Fegan, M., Hayward, C., Chakraborty, S., and Sly, L.I. 1999. Phylogenetic relationships among members of the Comamonadaceae, and description of Delftia acidovorans (den Dooren de Jong 1926 and Tamaoka et al. 1987) gen. nov., comb. nov. International Journal of Systematic and Evolutionary Microbiology **49**(2): 567-576.
Weygand-Durasevic, I., and Ibba, M. 2010. New roles for codon usage. Science **329**(5998): 1473-1474.
Wilson, D.N., and Nierhaus, K.H. 2007. The weird and wonderful world of bacterial ribosome regulation. Critical reviews in biochemistry and molecular biology **42**(3): 187-219. doi: 10.1080/10409230701360843.
Willems, A. 2014. The family Comamonadaceae. *In* The Prokaryotes. Springer. pp. 777-851.
Wolf, Y.I., Rogozin, I.B., Kondrashov, A.S., and Koonin, E.V. 2001. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome research **11**(3): 356-372.
Wolfenden, R.V., Cullis, P., and Southgate, C. 1979. Water, protein folding, and the genetic code. Science **206**(4418): 575-577.
Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution **24**(8): 1586-1591.
Zaslaver, A., Mayo, A., Ronen, M., and Alon, U. 2006. Optimal gene partition into operons correlates with gene functional order. Physical biology **3**(3): 183.
Zhang, B., Pan, X., Cox, S., Cobb, G., and Anderson, T. 2006. Evidence that miRNAs are different from other RNAs. Cellular and Molecular Life Sciences CMLS **63**(2): 246-254.
Zuker, M., and Stiegler, P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic acids research **9**(1): 133-148.

ANEXO: FIGURAS Y TABLAS

Tabla A1. Diferencia entre el contenido de GC total y en cada posición del codón entre los genes ortólogos y los todos los genes del genoma.

GC%	GC1%	GC2%	GC3%	Organismos
1.5	2	4.1	-1.6	Acidovorax avenae ATCC 19860
0.9	1.5	4.1	-3.1	Acidovorax citrulli AAC00 1
1.6	1.6	3.2	-0.1	Acidovorax ebreus TPSY
0.8	1	3	-1.8	Acidovorax JS42
1.5	1.4	3.1	0.1	Acidovorax_KKS102_uid176500
0.3	0.9	2.9	-2.8	Alicycliphilus denitrificans BC
0.4	1	3	-2.8	Alicycliphilus denitrificans K601
-2.8	-3	-0.3	-5	Comamonadaceae bacterium CR
0.3	0.1	2.7	-2	Comamonas testosteroni CNB 2
1.2	1.4	3.6	-1.3	Delftia_acidovorans_SPH_1_uid58703
1.7	1.7	3.6	-0.4	Delftia Cs1 4
0.4	0.2	2.5	-1.4	Polaromonas JS666
0.6	0.3	2.5	-0.9	Polaromonas naphthalenivorans CJ2
1.1	1.8	3.4	-2	Ramlibacter tataouinensis TTB310
0.5	-0.1	1.8	-0.3	Rhodoferax ferrireducens T118
1.2	1.3	3.4	-1	Variovorax paradoxus B4
1.4	1	3.1	0	Variovorax paradoxus EPS
1.1	1.2	3.4	-1.2	Variovorax paradoxus S110
-0.7	-1	2.2	-3.3	Verminephrobacter eiseniae EF01 2
2	2.1	4.5	-0.7	Acidovorax CF316
1.3	1	4.7	-1.9	Acidovorax delafieldii 2AN
1	0.9	3	-0.9	Acidovorax NO 1
1.9	1.5	3.4	0.7	Acidovorax radicis N35
0.2	-0.1	2.8	-2.2	Comamonas testosteroni ATCC 11996
0.1	-0.1	2.8	-2.5	Comamonas testosteroni KF 1
0.1	0	2.8	-2.4	Comamonas testosteroni S44
1.1	1.2	3.2	-1.2	Hydrogenophaga PBC
0.6	0.6	2.4	-1.1	Hylemonella gracilis ATCC 19624
1.2	0.9	3.2	-0.6	Polaromonas CF318
1.2	0.9	3.3	-0.7	Variovorax CF313
-0.7	-0.7	4.8	-6.1	Verminephrobacter At4
1.2	1.4	3.9	-1.5	Delftia sp. JD2

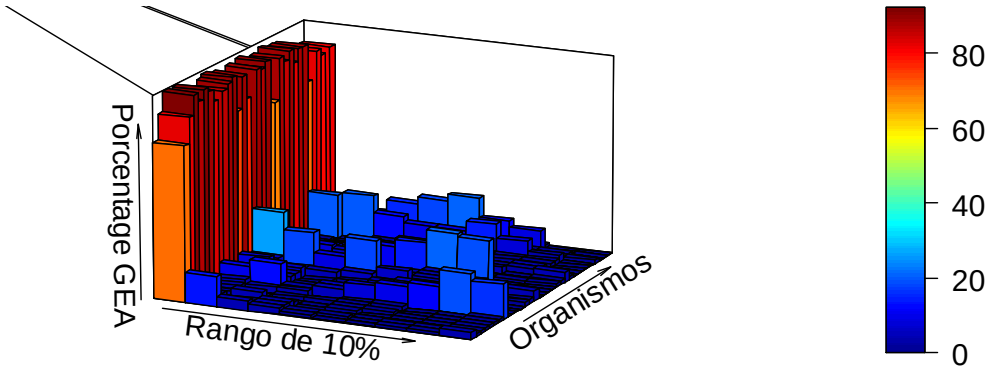

Figura A2. Se muestra el porcentaje de genes de alta expresión proteínas ribosomales y factores de elongación en los 32 organismos en los ejes del análisis de WCA que se correlacionan con el índice de expresión MELP. El eje que se correlacionan con el MELP se separó en 10 franjas.

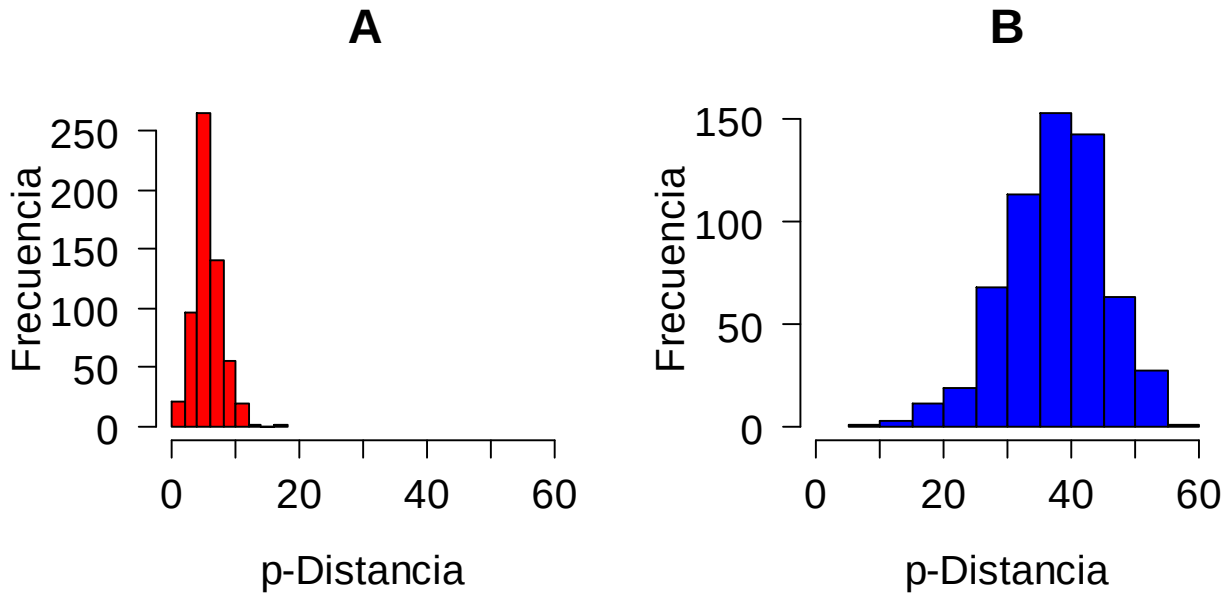

Figura A3. Se muestran las estimaciones de la p-distancia de las regiones conservadas y no conservadas. Histograma de la frecuencia de conservada (A) y regiones no conservadas (B) en las 601 proteínas ortólogas distribuidos de acuerdo a la p-distancia media. Las regiones conservadas se alinean en bloques con una longitud de al menos dos aminoácidos, que estén sin cambios en los 32 organismos y con un máximo de una posición contigua no conservada. El gráfico muestra que las regiones de las proteínas seleccionadas como conservada mediante el software Gblock se caracterizan por secuencias menos divergentes en comparación con las secuencias de las regiones no conservadas.

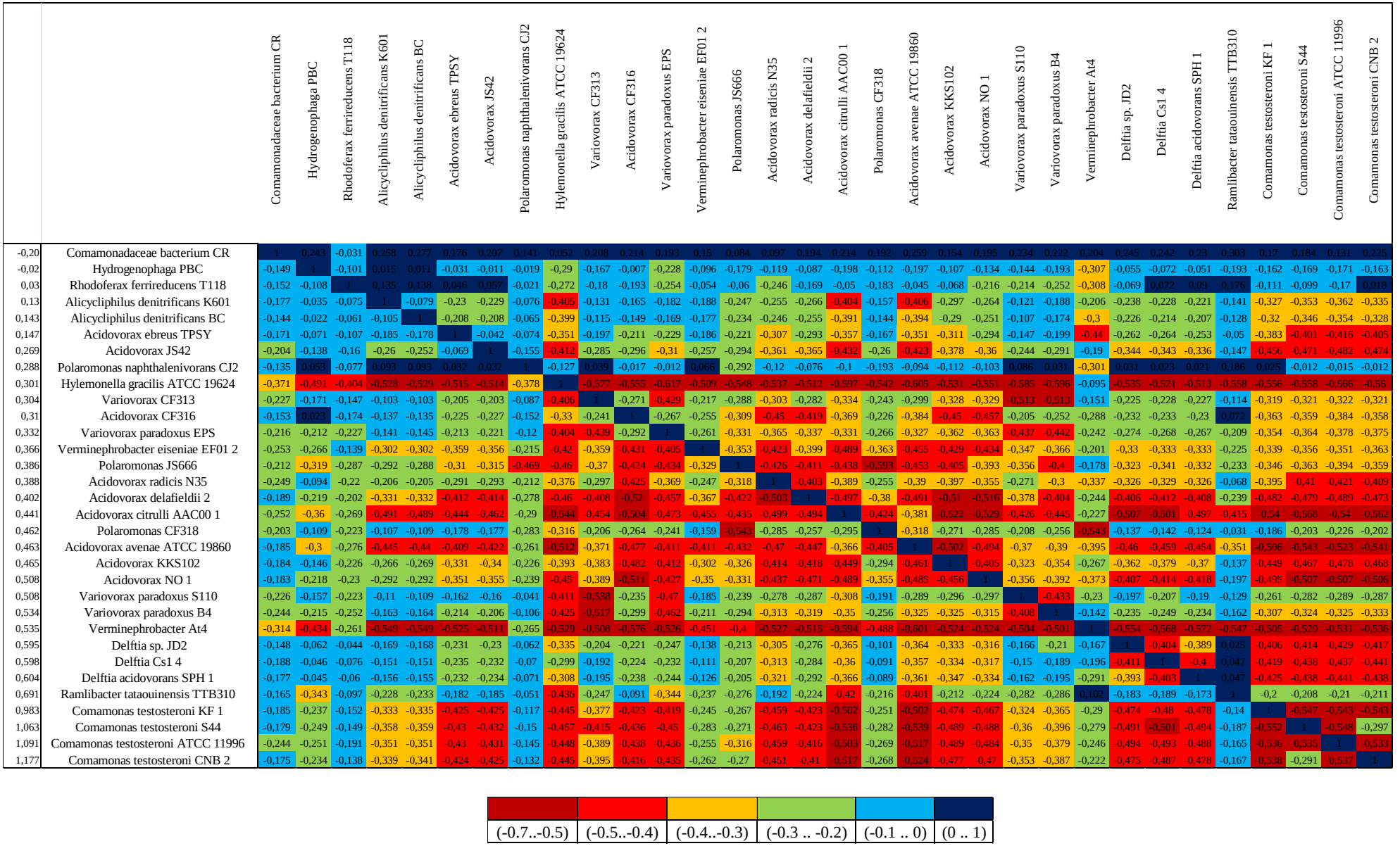

Figure A4. Codones preferidos y evitados utilizando a los genes ortólogos entre los GAEs en la familia *Comamonadaceae*. Se utilizó el estadístico $\chi 2$ de contingencia para poder determinar que codón es preferido o evitado en forma significativa por los genes GEA con respecto a los genes GBE ($p < 0,05$). En rojo y bordó estan representados los codones preferidos por los GEA ($p<0,01$ y $p<0,05$) respectivamente, mientras que en azul y violeta esta los codones evitados ($p<0,01$ y $p<0,05$) respectivamente.

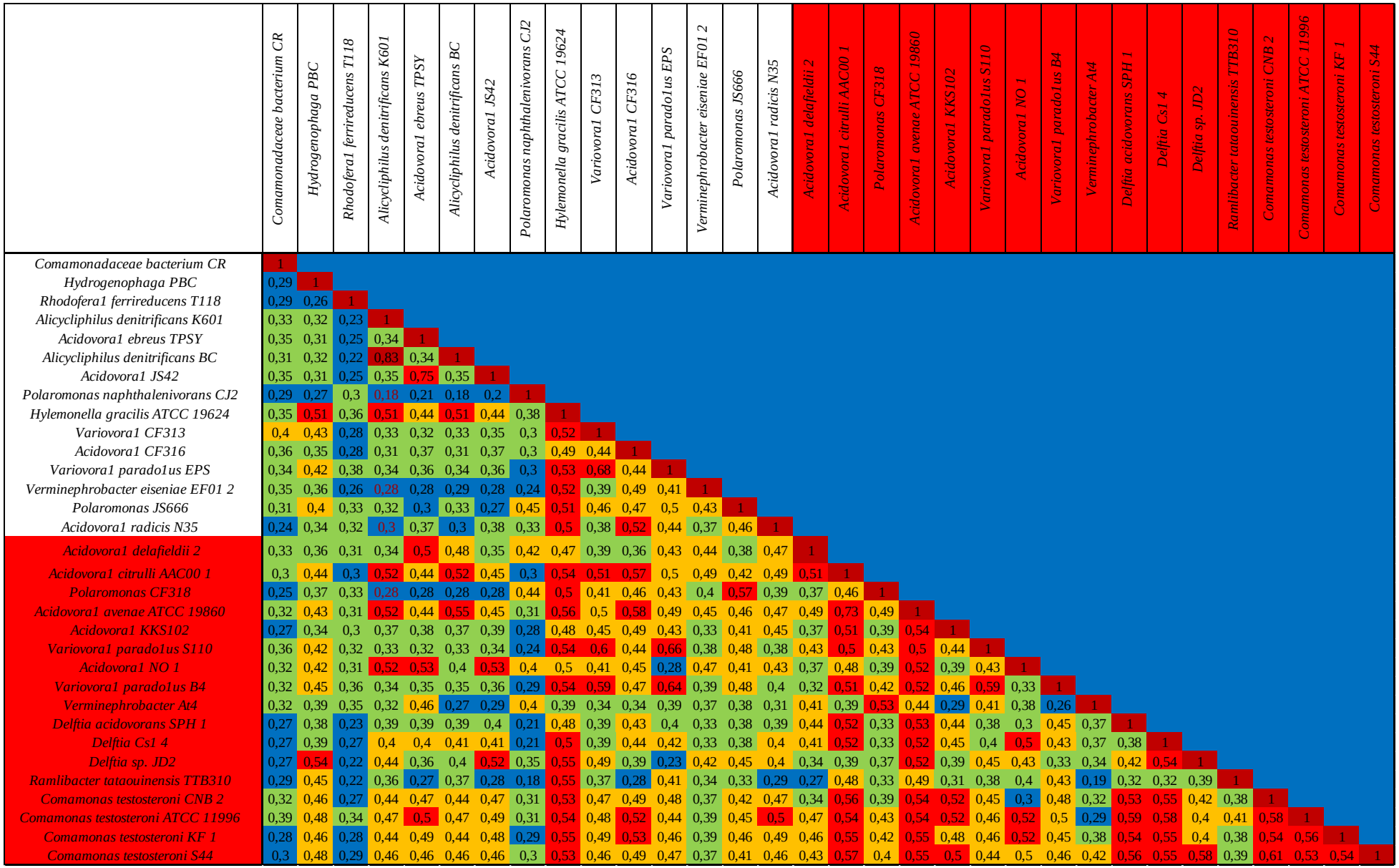
Comamonadaceae bacterium CR
Hydrogenophaga PBC
Rhodofera1 ferrireducens T118
Alicycliphilus denitrificans K601
Acidovora1 ebreus TPSY
Alicycliphilus denitrificans BC
Acidovora1 JS42
Polaromonas naphthalenivorans CJ2
Hylemonella gracilis ATCC 19624
Variovora1 CF313
Acidovora1 CF316
Variovora1 parado1us EPS
Verminephrobacter eiseniae EF01 2
Polaromonas JS666
Acidovora1 radicis N35
Acidovora1 delafieldii 2
Acidovora1 citrulli AAC00 1
Polaromonas CF318
Acidovora1 avenae ATCC 19860
Acidovora1 KKS102
Variovora1 parado1us S110
Acidovora1 NO 1
Variovora1 parado1us B4
Verminephrobacter At4
Delftia acidovorans SPH 1
Delftia Cs1 4
Delftia sp. JD2
Ramlibacter tataouinensis TTB310
Comamonas testosteroni CNB 2
Comamonas testosteroni ATCC 11996
Comamonas testosteroni KF 1
Comamonas testosteroni S44
(1..0.8)
(0.8..0.5)
(0.5..0.4)
(0.4 .. 0.3)
(0.3.. 0)

Figure A5. Se muestra la correlación entre las estimaciones de dS y dN utilizando a todos los organismos. Están coloreados con rojo los 17 de los cuales presentan valores de S mayores a 0.4 de S y mayores al 95% del límite de los valores de S tomando genes al azar.

Tabla A2. Porcentaje de GC entre los genes de las proteínas ribosomales y no ribosomales.

	Ribosomales	No Ribosomales	Significancia Wilcox.test
GC	0,61	0,66	P<0,05
GC1	0,63	0,67	P<0,05
GC2	0,42	0,47	P<0,05
GC3	0,76	0,86	P<0,05
G1	0,39	0,39	P>0,05
G2	0,18	0,20	P<0,05
G3	0,27	0,40	P<0,05
C1	0,24	0,28	P<0,05
C2	0,23	0,27	P<0,05
C3	0,49	0,46	P<0,05

Tabla A3. Uso relativo de aminoácidos de las proteínas ribosomales y de las proteínas no ribosomales.

Aminoácidos	Proteínas Ribosomales	Proteínas no Ribosomales	Significancia
Ile	0.054	0.042	P<0.05
Val	0.089	0.075	P<0.05
Leu	0.074	0.107	P<0.05
Phe	0.029	0.033	P<0.05
Cys	0.0038	0.009	P>0.05
Met	0.026	0.024	P<0.05
Ala	0.11	0.13	P<0.05
Gly	0.086	0.083	P<0.05
Thr	0.053	0.051	P<0.05

Ser	0.048	0.054	P<0.05
Trp	0.0046	0.015	P<0.05
Tyr	0.017	0.021	P<0.05
Pro	0.038	0.053	P<0.05
His	0.019	0.023	P<0.05
Asn	0.033	0.025	P<0.05
Asp	0.047	0.050	P<0.05
Gln	0.035	0.043	P<0.05
Glu	0.060	0.051	P<0.05
Lys	0.088	0.031	P<0.05
Arg	0.079	0.069	P<0.05

Tabla A4. Uso relativo de aminoácidos entre las regiones de las proteínas de membranas y entre proteínas de membranas y proteínas.

Aminoácidos	Proteínas TM	Proteínas no TM	Significancia	Regiones TM	Regiones NOTM	Significancia
Ile	0,046	0,041	P<0.05	0,070	0,032	P<0.05
Val	0,084	0,075	P<0.05	0,122	0,062	P<0.05
Leu	0,127	0,101	P<0.05	0,198	0,088	P<0.05
Phe	0,04	0,032	P<0.05	0,066	0,027	P<0.05
Cys	0,007	0,011	P<0.05	0,011	0,014	P>0.05
Met	0,028	0,024	P<0.05	0,034	0,022	P<0.05
Ala	0,137	0,128	P<0.05	0,163	0,131	P<0.05
Gly	0,087	0,083	P<0.05	0,1	0,083	P<0.05
Thr	0,051	0,052	P>0.05	0,045	0,055	P<0.05
Ser	0,057	0,054	P<0.05	0,046	0,067	P<0.05
Trp	0,02	0,013	P<0.05	0,032	0,019	P<0.05
Tyr	0,021	0,022	P<0.05	0,025	0,017	P<0.05
Pro	0,051	0,054	P<0.05	0,031	0,069	P<0.05
His	0,018	0,025	P<0.05	0,008	0,024	P<0.05
Asn	0,024	0,026	P<0.05	0,011	0,025	P<0.05
Asp	0,038	0,055	P<0.05	0,004	0,045	P<0.05
Gln	0,039	0,044	P<0.05	0,012	0,047	P<0.05
Glu	0,038	0,056	P<0.05	0,006	0,044	P<0.05
Lys	0,025	0,032	P<0.05	0,003	0,031	P<0.05
Arg	0,061	0,072	P<0.05	0,013	0,097	P<0.05

Tabla A5. Porcentaje de GC entre los genes de las regiones transmembranas y las regiones no no transmembranas.

	TM	No TM	Significancia Wilcox.test
GC	0,65	0,67	P<0,05
GC1	0,64	0,67	P<0,05
GC2	0,44	0,54	P<0,05
GC3	0,88	0,79	P<0,05
G1	0,4	0,37	P>0,05
G2	0,17	0,24	P<0,05
G3	0,45	0,38	P<0,05
C1	0,24	0,31	P<0,05
C2	0,27	0,3	P<0,05
C3	0,42	0,41	P<0,05