UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS
INSTITUTO DE INVESTIGACIÓN EN CIENCIAS BÁSICAS Y APLICADAS
CENTRO DE INVESTIGACIÓN EN DINÁMICA CELULAR

"Efecto del sesgo del contenido de GC genómico de organismos procariotas en las estructuras secundarias de sus proteínas"

TESIS

QUE PARA OBTENER EL GRADO DE

DOCTOR EN CIENCIAS

PRESENTA

DIANA BARCELÓ ANTEMATE

DIRECTOR DE TESIS
Dr. Enrique Merino Pérez

CUERNAVACA, MORELOS MARZO, 2023

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS
INSTITUTO DE INVESTIGACIÓN EN CIENCIAS BÁSICAS Y APLICADAS
CENTRO DE INVESTIGACIÓN EN DINÁMICA CELULAR

"Efecto del sesgo del contenido de GC genómico de organismos procariotas en las estructuras secundarias de sus proteínas"

TESIS

QUE PARA OBTENER EL GRADO DE

DOCTOR EN CIENCIAS

PRESENTA

DIANA BARCELÓ ANTEMATE

DIRECTOR DE TESIS
Dr. Enrique Merino Pérez

CUERNAVACA, MORELOS MARZO, 2023

LISTA DE JURADO

	Nombre (completo)	Adscripción	Línea de investigación	Correo Electrónico
DIRECTOR DE TESIS	Dr. Enrique Merino Pérez	IBT - UNAM	Bioinformática aplicada a las ciencias ómicas	enrique.merino@ibt.unam.mx
PRESIDENTE	Dra. Carmen Nina Pastor Colón	CIDC - UAEM	Estructura y Función de Macromoléculas	nina@uaem.mx
SECRETARIO	Dra. María del Rayo Sánchez Carbente	CEIB - UAEM	Biotecnología de plantas y microorganismos	maria.sanchez@uaem.mx
VOCAL	Dr. Ramón Antonio González García-Conde	CIDC - UAEM	Dinámica celular	rgonzalez@uaem.mx
VOCAL	Dra. Cinthia Ernestina Núñez López	IBT - UNAM	Microbiología molecular e industrial para la producción de biopolímeros	cinthia.nunez@ibt.unam.mx
VOCAL	Dra. Verónica Mercedes Narváez Padilla	CIDC - UAEM	Estructura y Función de Macromoléculas	vnarvaez@uaem.mx
SUPLENTE	Dra. Rosa María Gutiérrez Ríos	IBT - UNAM UNAM	Bioinformática de procesos de regulación en bacterias	rosa.gutierrez@ibt.unam.mx
SUPLENTE	Dr. Armando Hernández Mendoza	CIDC - UAEM	Estructura y Función de Macromoléculas	ahm@uaem.mx

LISTA DE PUBLICACIONES RELACIONADAS CON LA TESIS

Sueoka N. Correlation between Base Composition of Deoxyribonucleic Acid and Amino Acid Composition of Protein. *PROC N A S*. 1961;47:1141-1129.

Chou PY, Fasman GD. Prediction of Protein Conformation. *Biochemistry*. 1974;13(2):222-245. doi:10.1021/bi00699a002.

Chou PY, Fasman GD. Conformational parameters for amino acids in helical, Beta-sheet, and random coil region calculated from proteins. *Biochemetry*. 1974;13(2):211-222.

Singer GAC, Hickey DA. Nucleotide Bias Causes a Genomewide Bias in the Amino Acid Composition of Proteins. *Mol Biol Evol*. 2000;17(11):1581-1588.

Lightfield J, Fram NR, Ely B. Across bacterial phyla, distantly-related genomes with similar genomic GC content have similar patterns of amino acid usage. *PLoS One*. 2011;6(3). doi:10.1371/journal.pone.0017677.

Almpanis A, Swain M, Gatherer D, McEwan N. Correlation between bacterial G+C content, genome size and the G+C content of associated plasmids and bacteriophages. *Microb genomics*. 2018;4(4):e000168. doi:10.1099/mgen.0.000168.

AGRADECIMIENTOS

Agradezco a mi increíble director de tesis, el Dr. Enrique Merino Pérez, a quien admiro y estimo, por su guía, su tiempo (aún en fines de semana), sus enseñanzas, por su calidez humana, su paciencia y ánimos en cada etapa de mi Doctorado, porque no se me olvida la confianza que tuvo en mí antes de empezar el posgrado. ¡Sin duda un digno ejemplo a seguir!

Al Dr. Ramón González García-Conde por sus preguntas en cada tutorial, por sus valiosas sugerencias, por la rapidez en la que contesta los correos y la flexibilidad de apoyo en muchas cuestiones académicas. Valoro el tiempo que me ha compartido, sus conocimientos y admiro al investigador de talla completa que es.

A la Dra. Sonia Dávila Ramos por toda su amabilidad y profesionalismo, por las importantes recomendaciones en mi trayectoria académica. A quien agradezco mucho su tiempo e intervenciones tan puntuales y acertadas. Por ser una investigadora joven que me motiva, me inspira y que también admiro mucho.

A la Dra. Rosa María Gutiérrez Ríos por sus valiosos y acertados comentarios en cada seminario de grupo, por su apoyo en la revisión de estilo del artículo científico, por ser una investigadora a la cual admiro muchísimo.

A la M. en C. María Luisa Tabche por su apoyo al conseguir reactivos, por las enseñanzas en el uso de equipos y técnicas de laboratorio, por la fuerza que proyecta como mujer, por su calidez de “mamá” en el laboratorio, por su sinceridad y buenos consejos, por ser alguien confiable.

A Ricardo Ciria por sus comentarios ortográficos, por la buena disposición en enseñarnos y explicarnos las herramientas bioinformáticas y por la amabilidad en la que me recibió al llegar al laboratorio. Gracias por todo y por el termo rosa.

A mis mejores amigas, Nataly Morales Galeana y Nori Castañeda Gómez, por hacer muy divertida mi llegada y estancia en el laboratorio. Por sus aportaciones intelectuales, por su comprensión, por el tiempo que han tenido conmigo, por las aventuras y por la amistad tan bonita que hemos formado.

A todos los del laboratorio, Mariela Serrano, Maricela Carrera, Raúl Noguez, MariCarmen Sanchéz, Jannette Huerta, Brandon, Lizzeth Soto, que de alguna u otra forma han aportado perspectiva con diferentes puntos de vista, por la buena disposición que les caracteriza y por los momentos que nos han unido como buenos hermanitos de laboratorio.

Al personal de la UAEM, especialmente al Posgrado en Ciencias, a las secretarias Dulce Verónica y Cris Aranda, así como a Esmeralda González, quienes me han apoyado y orientado de la manera más amable posible en todos estos años del Doctorado.

A mis amigas de ayer y siempre, Yormery y Annely que aún en la distancia me dan ánimos se seguir y no rendirme en el camino de la investigación.

Este trabajo se desarrolló en el departamento de Microbiología molecular del Instituto de Biotecnología-UNAM, bajo la dirección del Dr Enrique Merino Pérez. Así mismo, como estudiante del Posgrado en Ciencias de la Universidad Autónoma del Estado de Morelos (UAEM), agradezco el apoyo económico al Consejo Nacional de Ciencia y Tecnología (CONACYT) por la beca con número de CVU 563693.

DEDICATORIA

Al amor de mi vida, mi esposo Fernando Fontove Herrera, por todo lo que eres y lo que me ofreces, quien no sólo me ha apoyado sentimentalmente, sino también psicológica y académicamente… ¡Me llena de satisfacción intelectual saber que vamos a publicar juntos un artículo científico!
Con mucho amor, tu esposa.

A mis padres, Lorenzo Barceló Aguilar y Josefa Antemate Chigo, por el invaluable apoyo que me han brindado en este camino de la investigación, sus ánimos, sus consejos en momentos difíciles o no tan claros.
Con amor, la más pequeña de sus hijas.

A mis hermanas, ¡simplemente gracias! Las amo a los dos con su característica forma de ser de cada una. Las llevo siempre en mi corazón.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE TABLAS COMPLEMENTARIAS

ÍNDICE DE FIGURAS COMPLEMENTARIAS

RESUMEN

Una de las principales características de los genomas procariotas es la proporción en que las bases Guanina-Citosina se utilizan en sus secuencias genómicas. Esto se conoce como el contenido de GC genómico y varía ampliamente, desde valores tan pequeños de 13% hasta valores tan grandes de 74%. Diversos estudios en microorganismos han demostrado que el contenido de GC genómico influye en la composición de aminoácidos de sus proteínas a nivel de su estructura primaria. Desde Sueoka[1], se ha observado que el sesgo que impone el contenido de GC genómico es particularmente importante en la composición de aminoácidos codificados por codones altos y bajos en contenido de GC, como Ala, Gly y Pro, y como Lys, Asn e Ile, respectivamente. En nuestro estudio, ampliamos estos resultados considerando el efecto del contenido de GC genómico en las estructuras secundarias de las proteínas. El presente estudio se realizó a través de un análisis bioinformático cuyos resultados se presentan y discuten en cuatro partes. Primero, a partir de un primer grupo de 192 procariotas con contenido de GC genómico del 20% al 74% se observó la distribución filogenética, obteniendo características propias de cada filo estudiado. Segundo, con los genomas y proteomas del primer grupo de estudio, y basados en los parámetros conformacionales de Chou y Fasman[2,3], encontramos que la tendencia de un aminoácido a formar parte de una estructura secundaria de proteínas no es absoluto, sino que varía según el contenido de GC genómico. Tercero, se identificó que la composición de las estructuras secundarias de los proteomas del primer grupo de estudio varía en relación con el contenido de GC genómico; esto es, a medida que aumenta el contenido genómico de GC los lazos aumentan, mientras que las hélices alfa y las hojas beta disminuyen. Cuarto, a partir de un segundo grupo de 1,544 procariotas se realizó un estudio donde se impusieron criterios estrictos de agrupación de proteínas ortólogas y descubrimos que, para algunos COGs el contenido de GC de los genes sesga la composición de las estructuras secundarias de las proteínas que codifican.

ABSTRACT

One of the main characteristics of prokaryotic genomes is the ratio in which Guanine-Cytosine bases are used in their genomic sequences. This is known as the genomic GC content and varies widely, from values as small as 13% to values as large as 74%. Several studies in microorganisms have demonstrated that the genomic GC content influences the amino acid composition of their proteins at their primary structural level. Since Sueoka[1], it has been observed that the bias imposed by genomic GC content is particularly important in the composition of amino acids encoded by codons high and low in GC, such as Ala, Gly, and Pro, and such as Lys, Asn, and Ile respectively. In our study, we extend these results by considering the effect of the genomic GC content on the secondary structure of proteins. The present study was conducted through a bioinformatics analysis whose results are presented and discussed in four parts. First, from the first group of 192 prokaryotes with a genomic GC content of 20 to 74%, the phylogenetic distribution was observed, obtaining characteristics of each phylum studied. Second, with the genomes and proteome of the first study group and based on studies of Chou and Fasman's conformational parameters[2,3], we found that the tendency of an amino acid to form part of a secondary structure of proteins is not absolute but varies according to the genomic GC content. Third, it was identified that the composition of the secondary structures of the proteomes of the first study group varies in relation to the genomic GC content; that is, as the genomic GC content increases, coils increase, while alpha-helices and beta-sheets decrease. Four, from a second group of 1,544 prokaryotes, we did a study where were imposed strict criteria of orthologous protein clustering and we found that, for some COGs the GC content of genes biases the composition of secondary structures of the proteins for which they code.

1. INTRODUCCIÓN

El contenido de GC (composición de Guanina-Citosina) es un parámetro clave de la variación genómica en todos los organismos[4]. Es bien conocido que el contenido de GC genómico en procariotas varía marcadamente. Los estudios demuestran que la variación en el contenido de GC de los genomas de microorganismos actualmente secuenciados va del 13% al 75%, aproximadamente[5,6].

Las causas de la gran diferencia en el contenido de GC genómico de procariotas aún no se conocen con certeza, pero se han correlacionado con el tamaño del genoma[7,8], el estilo de vida (simbiontes o de vida libre)[9,10], los hábitats ambientales (como ambientes extremos)[11–13], las condiciones ambientales (como temperatura y pH,)[4,14,15], la relación filogenética[16], la presión mutacional[17–19], entre otras variables.

Ha sido ampliamente documentado que la variación del contenido de GC genómico es uno de los principales contribuyentes de la arquitectura proteómica en un organismo[1,12,20–23], ya que impacta directamente en los aminoácidos de sus proteínas a nivel estructural primario. Los aminoácidos principalmente afectados son aquellos codificados por codones altos y bajos en contenido de GC.

Un aspecto aun no considerado, y que es la base de este proyecto de investigación, es el estudio de la relación de la variación del contenido de GC genómico y la composición de las estructuras secundarias de las proteínas.

Los aminoácidos tienden a ser parte de los diferentes elementos de estructuras secundarias de las proteínas[24,25]. Esto ha sido determinado en base a la frecuencia con la que los aminoácidos están presentes en dichos elementos, siendo consistentes con sus propiedades fisicoquímicas y confiriendo estabilidad[2,3,26]. Dicha tendencia ha sido establecida con valores numéricos conocidos como parámetros conformacionales y pueden usarse para clasificar a los aminoácidos como formadores, indiferentes e interruptores de las estructuras secundarias de las proteínas[2,3].

Nuestro estudio muestra por primera vez el impacto del contenido de GC genómico en las estructuras secundarias de las proteínas de organismos procariotas. Los resultados se presentan y discuten en diferentes niveles:

- Taxonómico, analizando la distribución del contenido de GC genómico en diferentes clados filogenéticos.

- Proteómico, evaluando el impacto del contenido de GC genómico sobre la tendencia de los aminoácidos a formar parte de los elementos de estructura secundaria de las proteínas (parámetros conformacionales);

- Proteómico, estimando el impacto del contenido de GC genómico sobre las frecuencias relativas de las estructuras secundarias de las proteínas;

- Evolutivo-funcional, analizando el efecto del contenido de GC de los genes ortólogos en las frecuencias relativas de las diferentes estructuras secundarias de las proteínas a las que codifican.

2. ANTECEDENTES

Durante muchas décadas ha resultado atractivo conocer cómo la variación en el contenido de GC de los genomas de microorganismos impacta en las proteínas que conforman al proteoma. En el presente proyecto mencionaremos los principales trabajos que se han realizado en relación con el contenido de GC genómico y la composición de las proteínas a nivel estructural primario, los problemas que han resuelto, sus aportaciones, así como la relación que tiene con otras características biológicas.

Otro tema que también se abordará es la estructura secundaria de las proteínas, en la cual los aminoácidos forman a estos elementos. Detallaremos los principales

trabajos que han sido referencia en la descripción de la preferencia de los aminoácidos a formar parte de los diferentes tipos de estructuras secundarias de las proteínas.

2.1 Influencia del contenido de GC genómico en el uso de codones de los aminoácidos de las proteínas

Debido a que el código genético es altamente degenerado[27,28], 18 de los 20 aminoácidos son codificados por más de un codón. En consecuencia, hay aminoácidos como Ala, Gly Pro y Arg que son codificados por codones altos en GC, mientras que otros como Tyr, Lys, Ans e Ile son codificados por codones altos en AT[19–21].

Es bien sabido que no todos los codones y aminoácidos son igualmente utilizados por los organismos. Diversos estudios demuestran que existe una preferencia por el uso de codones que se correlaciona linealmente con el contenido de GC genómico a través de los filos y con el contenido relativo de aminoácidos de sus correspondientes proteomas[13,16,29].

En relación a lo anterior, para determinar el impacto de la variación del contenido de GC genómico en el uso de codones de los aminoácidos[30], se ha estudiado el contenido de GC de cada posición de los codones[16,29]. Estudios previos demuestran consistencia en el hecho de que GC de la tercera posición de un codón aminoacídico incrementa rápidamente con el incremento del contenido de GC genómico. Esto es que el GC de la tercera posición del codón cambia más rápido en comparación con el GC de la primera y segunda posición del codón y no depende directamente del aminoácido que codifica[16,17,31].

2.2 Estudios del sesgo del GC genómico en la frecuencia de aminoácidos de las proteínas a nivel estructural primario

El genoma y su contenido de GC ha sido centro de atención para explicar la influencia directa sobre el contenido de aminoácidos de las proteínas (proteoma) de organismos representantes en los tres dominios de la vida: Bacteria, Arquea y Eucariota[12]. Ha sido ampliamente abordado que los aminoácidos con codones altos o bajos en GC son más sensibles a la variación del contenido de GC genómico[1,16,20,22,31]. La siguiente Tabla 1 clasifica a los 20 aminoácidos en tres grupos: con alto, neutro y bajo contenido de GC en sus codones.

Tabla 1. Clasificación de los aminoácidos del Código Genético con respecto al contenido de GC de sus codones. Los aminoácidos (referidos en código de tres letras), los codones y el contenido de GC de los codones son clasificados en tres grupos.

	Aminoácido	Secuencia de codones						%GC
Codones con GC bajo	Ile	ATT	ATA	ATC				11.1
	Asn	AAT	AAC					16.7
	Lys	AAA	AAG					16.7
	Tyr	TAT	TAC					16.7
	Phe	TTT	TTC					16.7
	Met	ATG						33.3
	Leu	TTA	TTG	CTT	CTA	CTC	CTG	38.9
Codones con GC neutro	Ser	TCT	TCA	TCC	TCG	AGT	AGC	50.0
	Glu	GAA	GAG					50.0
	Cys	TGT	TGC					50.0
	Gln	CAA	CAG					50.0
	Thr	ACT	ACA	ACC	ACG			50.0
	Asp	GAT	GAC					50.0
	His	CAT	CAC					50.0
	Val	GTT	GTA	GTC	GTG			50.0
Codones con GC alto	Trp	TGG						66.7
	Arg	AGA	AGG	CGT	CGA	CGC	CGG	72.2
	Pro	CCT	CCA	CCC	CCG			83.3
	Gly	GGT	GGA	GGC	GGG			83.3
	Ala	GCT	GCA	GCC	GCG			83.3

Los nucleótidos G y C son resaltados en rojo. El contenido de GC de los codones son expresados como porcentaje en la última columna. Los 20 aminoácidos estándar fueron clasificados según el número de nucleótidos G o C, dividido por el número de nucleótidos en sus respectivos codones.

Desde 1961 Sueoka fue pionero en realizar estudios sobre la influencia del GC a partir de secuencias genómicas en los aminoácidos. Sueoka encontró una correlación positiva en algunos aminoácidos como Ala, Arg, Gly y Pro, y una correlación negativa en Ile, Lys y Asn con respecto al incremento del contenido de GC genómico de los organismos de estudio[1].

Otro trabajo que ha validado el descubrimiento de Sueoka es el estudio de Lobry en 1997. Este estudio se basó en ver cómo influía los diferentes contenidos de GC genómico de 59 especies microbianas (bacterias y arqueas) en dos grupos de proteínas: las proteínas integrales de membranas y las proteínas periféricas. Lobry encontró, en ambos grupos de proteínas, que cuando el GC genómico incrementaba, la frecuencia de los aminoácidos Ala, Gly, Pro y Arg de los proteomas tendían a aumentar, mientras que las de Tyr, Lys, Ans e Ile tendían a disminuir[22].

El estudio de Singer y Hickey en el 2000 abarcó diferentes contenidos de GC a partir de secuencias genómicas bacterianas, de levadura y de protozoarios y registró la influencia que ejercía en la composición de aminoácidos clasificados en dos grupos: aquellos con codones ricos en GC (Gly, Ala, Arg, Pro) y con codones ricos en AT (Phe, Tyr, Met, Ile, Asn, Lys), encontrando que, a mayor contenido de GC genómico la presencia de aminoácidos con codones ricos en GC incrementaba, observando un comportamiento contrario en el grupo con codones ricos en AT[20].

Otro estudio importante por mencionar es de Lightfield y colaboradores en el 2011. Su grupo de estudio fueron especies representativas de cinco filos bacterianos con diferente contenido de GC genómico. Entre los resultados observados fue que, cuando aumentaba el contenido de GC genómico, aumentaba el uso de codones sinónimos de un aminoácido que empiezan con G ó C (e.g. Arg, es codificada con cuatro codones que empiezan con C: CGU, CGC, CGA, CGG) y disminuía con aquellos que empiezan con A (e.g. Arg, es codificada con dos codones que empiezan con A: AGA, AGG)[16].

2.3 Relación del contenido de GC genómico y la longitud del genoma

Una característica abordada en varios estudios, consecuente de la variación del contenido de GC genómico, es el tamaño de genoma. El genoma procariota varía considerablemente y puede ser tan pequeño[5,32] y simple como el genoma de 109 kilobases de *Candidatus Nasuia deltocephalinicola*[8,33], o tan grande y complejo como el genoma de 16 megabases de *Minicystis rosea*[8,34].

Uno de los trabajos más actualizados en estudiar la relación del contenido de GC genómico bacteriano con el tamaño del genoma y sus plásmidos fue el de Almpanis y colaboradores, en el 2018. Ellos encontraron que los genomas más largos tienden a tener contenidos de GC más grandes. Un patrón similar adoptaban sus plásmidos: plásmidos más grandes tuvieron contenidos de GC más grandes, además encontraron una correlación positiva entre el contenido de GC de los plásmidos y el contenido de GC genómico de su bacteria hospedera[8].

2.4 Propuestas para explicar las variaciones en el contenido de GC genómico

Para explicar la variación en el contenido de GC genómico se han formulado varias teorías e hipótesis que resultan importante de mencionar. Los factores que han influenciado esta variación han sido debatidos desde hace 60 años[4,8,15,18,35–39,40–42]. Las principales teorías que se establecen son la teoría de mutación direccional y la teoría de la fuerza selectiva, las cuales se explicarán a continuación.

La teoría de la presión de mutación direccional, propuesta en 1962 por Sueoka, se basa en que el efecto de la mutación sobre el genoma no es azaroso y tiene direccionalidad hacia el más alto o el más bajo contenido de GC del genoma. Esta teoría también considera el aspecto filogenético explicando la amplia variación en el contenido

de GC genómico entre diferentes bacterias y su pequeña heterogeneidad dentro de especies bacterianas individuales[18,38].

En contraste, la teoría de la fuerza selectiva nos habla del rol de la selección para determinar la composición GC del genoma. Debido a que los genomas bacterianos están compuestos principalmente por genes que codifican proteínas, la fuerza selectiva actúa sobre cada gen para aumentar su contenido de GC y así influye acumulativamente en la composición general de bases genómicas[39].

En relación a las teorías antes mencionadas, el trabajo de Wu y colaboradores[4], resume las siguientes hipótesis para explicar las variaciones del contenido de GC genómico en procariotas:

- De resistencia a UV: Alto contenido de GC genómico ofrece una ventaja selectiva a organismos que viven en ambientes que son susceptibles a la luz solar directa e intensa[41]. Ejemplos de estos organismos son representantes de los géneros Actinoplanes, Cellulomonas, Streptomyces, Micromonospora, entre otros.

- De adaptación térmica: Organismos termófilos demuestran una tendencia a tener alto contenido de GC genómico debido a que la termo-estabilidad y termo-labilidad de los aminoácidos son reguladas por codones ricos y bajos en contenido de GC, respectivamente[7,15]*. Thermus thermophilus* como caso de estudio.

- De fijación de Nitrógeno: Hay contenido de GC genómico significativamente más alto en los miembros del género que fijan nitrógeno que en aquellos que no tienen la habilidad de fijarlo[42]. Ejemplo de las bacterias fijadoras de nitrógeno con más alto GC genómico son del género Aquaspirilum: *A. fasciculus*, *A. itersonii*, *A. magnetotacticum*, *A. peregrinum*; como del género Vibrio: *V. diazotrophicus*, *V. natriegens*, *V, pelagius*.

- De requerimiento de Oxígeno: Procariotas aerobios presentan un incremento significativo en el contenido de GC genómico en relación con los organismos anaerobios[37]. Un ejemplo de ellos son las bacterias pertenecientes al género Microbacterium: *M. imperiale*, *M. lacticum* y *M. laevaniformans*; del género Micrococcus, Halobacterium, Aquaspirillum, entre otros.

- De la DNA polimerasa III: La variación del contenido de GC es gobernada por los mecanismos de replicación y reparación del DNA. De acuerdo a la combinación dimérica de las subunidades alfa de la DNA pol III, el contenido de GC de los genomas de eubacterias son divididos en tres grupos con distinto espectro de contenido de GC, *dnaE1* (espectro completo), *dnaE2/E1* (alto GC), y *polC/dnaE3* (bajo GC)[4,35]. Como caso de estudio, representantes del género Deinococcus y Thermus fueron evaluadas.

2.5 Las proteínas y su segundo nivel de organización estructural

El conjunto de proteínas codificadas por el genoma constituye el proteoma de un organismo. A su vez, las proteínas tienen organización de carácter jerarquizado: estructura primaria, secundaria, terciaria y cuaternaria[43].

Desde 1951, Paulin y Corey (in Pirovano & Heringa, 2010) sugirieron la existencia de conformaciones regulares dependientes de la secuencia de aminoácidos en las proteínas[25]. Sus estudios apuntaban a dos conformaciones particularmente estables que dan lugar a patrones estructurales repetitivos: hélice alfa y hoja beta. Más tarde, en algunas partes de la proteína, aparecieron pliegues menos regulares. Esta tercera clase de regiones menos estructuradas se conoce comúnmente como lazo[25,44] y junto con las hélices alfa y hojas beta conforman a los elementos de estructura secundaria de una proteína.

La importancia de los aminoácidos y su secuencia radica en que han sido la esencia para predecir las estructuras secundarias de las proteínas[24], donde llegan a

tener una eficiencia de hasta 80%. Del mismo modo, la distribución de hélices alfa, hojas beta y lazos han sido base en la estructura de las proteínas para clasificarlas según la forma y predecir posible función[25].

2.6 Aminoácidos formadores e interruptores de las estructuras secundarias de las proteínas

Trabajos como el de Chou y Fasman, así como el de Lewis son considerados pioneros por determinar que los aminoácidos tienden a formar o interrumpir los diferentes elementos de estructura secundaria en las proteínas[2,3,45].

En 1974 Chou y Fasman observaron la frecuencia que ciertos aminoácidos tenían con respecto a las hélices alfa, hojas beta, así como de los lazos. Ellos encontraron que hay aminoácidos formadores, indiferentes e interruptores y esto dependía del valor conocido como Parámetro Conformacional (PC)[2,3]. Un ejemplo es el aminoácido Pro, considerado un fuerte interruptor en hélices alfa (PC =0.59) y hojas beta (PC = 0.62), y de alta ocurrencia en lazos (PC = 1.45).

Otro trabajo apoyado en las investigaciones de Chou y Fasman, fue el de Argos y Palau. Este estudio habla, en términos porcentuales, de la composición de los aminoácidos en hélices alfa y hojas beta, además mencionan a los aminoácidos encontrados en los lazos, previos y posteriores de tales estructuras[26].

Los parámetros conformacionales de los aminoácidos han sido descritos e idealizados como valores invariantes en todos los organismos. Además han sido referidos en muchas estudios de predicción de estructura secundaria de las proteínas[46].

3. JUSTIFICACIÓN

Nuestro estudio se centra en el análisis de microorganismos debido a su diversidad, abundancia y su enorme importancia básica y aplicada[53]. La razón de estudiar el contenido de GC genómico en procariotas es por el tamaño pequeño de sus genomas, amplia variación del contenido de GC en genoma y el progresivo incremento y accesibilidad de los genomas y proteomas. Adicionalmente, distintas bases de datos se siguen enriqueciendo con una gran cantidad de información nucleotídica, aminoacídica y estructural[47–49], así como han surgido varias herramientas de apoyo para su análisis[50–53].

Un aspecto importante que nuestro proyecto trata, y del cual no hay precedente, es la influencia del contenido de GC genómico a nivel estructural secundario de las proteínas. En este sentido, nuestros resultados aportan nuevas evidencias y robustecen el hecho de que el contenido de GC genómico es una característica muy importante y esencial de los organismos, ya que contribuye fuertemente en la arquitectura proteómica, influyendo no solo en la estructura primaria (aminoácidos) sino en las estructuras secundarias (hélice alfa, hoja beta y lazo) de las proteínas.

Por tanto, aún falta un largo camino para comprender la importancia de conocer hasta dónde el contenido de GC genómico puede intervenir, y de qué forma, en ventajas biológicas y/o evolutivas es un camino largo por recorrer. Este proyecto de investigación avanzará una parte del camino, y al mismo tiempo, abrirá paso a nuevas preguntas de investigación.

4. HIPÓTESIS

El contenido de GC genómico de los organismos procariotas impone un sesgo en los elementos de estructura secundaria de sus correspondientes proteínas. Este sesgo se verá reflejado en:

i. Variación en la tendencia de un aminoácido a formar parte de una estructura secundaria;

ii. Cambios en los valores de los parámetros conformacionales de aminoácidos de las estructuras secundarias de las proteínas del proteoma;

iii. La composición de los elementos de estructura secundaria de las proteínas del proteoma: hélice alfa, hoja beta y lazo;

iv. El contenido de las estructuras secundarias de las proteínas ortólogas.

5. OBJETIVOS

5.1 OBJETIVO GENERAL

Evaluar, mediante un análisis *in silico*, el posible sesgo que introduce el contenido de GC genómico sobre las estructuras secundarias de las proteínas en bacterias y arqueas.

5.2 OBJETIVOS ESPECÍFICOS

a) Determinar el contenido de GC genómico a partir de organismos cuya secuencia genómica haya sido totalmente secuenciada y sean representativos únicos a nivel filogenético de especie.

b) Establecer la posible relación del contenido de GC genómico de los organismos de estudio con su origen filogenético.

c) Analizar la frecuencia relativa de los aminoácidos en las estructuras secundarias de las proteínas de los organismos de estudio en relación con el contenido de GC genómico.

d) Determinar los parámetros conformacionales de los aminoácidos de todas las proteínas de los organismos de estudio en relación con la variación del contenido de GC genómico.

e) Evaluar la relación de la frecuencia relativa de las estructuras secundarias (hélice alfa, hoja beta y lazo) de las proteínas de los organismos de estudio con el contenido de GC genómico.

f) Generar grupos de proteínas ortólogas (COGs) de procariotas representativos a nivel género y determinar la frecuencia relativa de las estructuras secundarias (hélice alfa, hoja beta y lazo) de sus proteínas ortólogas en relación con el contenido de GC de los genes que las codifican.

g) Identificar qué partes de las estructuras secundarias de las proteínas de cada COG son las más afectadas por el contenido de GC de los genes que las codifican.

6. MATERIALES Y MÉTODOS

6.1 Grupo de datos genómicos y proteómicos

Los datos de secuencias genómicas y proteómicas se dividen en dos grandes grupos de estudio:

- El primero consiste en un conjunto de 4,655 especies procariotas no redundantes extraídas de la base de datos KEGG GENOME en mayo de 2022 (Anexo 1). Para evitar la sobrerrepresentación de un contenido de GC dado, se seleccionaron un total de 192 organismos con contenidos de GC genómico que fueran lo más diferente posible, oscilando entre el 20% a 74% y abarcó los dominios Bacteria y Arquea. La secuencia de aminoácidos, usados para el estudio de estructuras secundarias de las proteínas de los organismos seleccionados, también se obtuvo de la misma base de datos.

- El segundo grupo fue usado para nuestro análisis de proteínas ortólogas. Un conjunto de 1,544 procariotas representativos no redundantes a nivel género también fue obtenido de la base de KEGG GENOME. Asimismo, este segundo grupo fue caracterizado en términos de su contenido de GC y la estructura secundaria de sus proteínas.

6.2 Clasificación taxonómica

La asignación taxonómica de nuestros grupos de organismos de estudio fue llevada a cabo de acuerdo con la base de datos KEGG GENOME[54] (https://www.genome.jp/kegg/genome/).

6.3 Evaluación del contenido de GC genómico

El propósito de este estudio es el análisis de las estructuras primarias y secundarias de las proteínas en relación con el contenido de GC del genoma en el que están codificadas. Teniendo en cuenta eso, para las evaluaciones del contenido de GC genómico solo se consideraron las regiones genómicas correspondientes a los genes que codifican proteínas. Las coordenadas genómicas de los genes que codifican proteínas se tomaron de la base de datos KEGG GENES[54] (https://www.genome.jp/kegg/genes.html).

Se calculó el contenido de GC genómico de la región codificante para cada uno de los 192 organismos seleccionados usando la siguiente fórmula:

$$GC_G = \frac{\sum_{g \in P(G)} GC(g)}{\sum_{g \in G} |g|}$$

donde *G* es el genoma de interés, *P(G)* son los genes que codifican proteínas de *G* (el proteoma de *G*), *g* es un gen dado que codifica a una proteína, |*g*| es su longitud, y *GC(g)* es el número de nucleótidos G o C en *g*.

6.4 Clasificación de los aminoácidos de acuerdo con el contenido de GC de sus codones

Se clasificaron los 20 aminoácidos del Código Genético según el contenido de GC de sus codones en tres grupos (Tabla 1). El primer grupo tiene 5 aminoácidos con alto contenido de GC en sus codones: Ala, Gly, Pro, Arg y Trp; el segundo tiene 8 aminoácidos con contenido neutro de GC en sus codones: Val, His, Asp, Thr, Gln, Cys, Glu y Ser; y el tercer grupo tiene 7 aminoácidos con bajo contenido de GC en sus codones: Leu, Met, Phe, Tyr, Lys, Asn e Ile.

6.5 Evaluación de la frecuencia de aminoácidos en los proteomas

Para cada proteoma de nuestros 192 organismos de estudio se evaluó la frecuencia relativa de cada uno de los 20 aminoácidos usando la siguiente fórmula:

$$aa^G = \frac{\sum_{p \in P(G)} |p(aa)|}{|P(G)|},$$

donde *G* es el genoma de interés, *P(G)* es su proteoma, *aa* es un aminoácido dado, *p* es una proteína, |*p(aa)*| es el número de veces que el aminoácido *aa* aparece en la proteína *p*, y |*P(G)*| es el número total de aminoácidos en el proteoma *P* de *G*.

6.6 Predicción de las estructuras secundarias de las proteínas

Los elementos de estructura secundaria de las proteínas, hélices alfa, hojas beta y lazos, se predijo usando PSSPRED[55] versión 4 (https://seq2fun.dcmb.med.umich.edu//PSSpred/).

6.7 Evaluación de la frecuencia de aminoácidos en las estructuras secundarias de los proteomas

Se analizó el contenido de los aminoácidos en las diferentes estructuras secundarias de las proteínas de nuestro primer grupo de estudio, compuesto por 192 procariotas, usando las siguientes fórmulas:

$$aa_\alpha^G = \frac{\sum_{p \in P(G)} |\alpha(p, aa)|}{\sum_{p \in P(G)} |\alpha(p)|},$$

$$aa_\beta^G = \frac{\sum_{p \in P(G)} |\beta(p, aa)|}{\sum_{p \in P(G)} |\beta(p)|},$$

$$aa_{\gamma}^{G} = \frac{\sum_{p \in P(G)} |\gamma(p, aa)|}{\sum_{p \in P(G)} |\gamma(p)|},$$

donde *G* es el genoma de interés, *P(G)* su proteoma, *aa* es un aminoácido dado, *p* es una proteína, |*ss(p,aa)*| es el número de veces que *aa* aparece en la estructura secundaria *ss* (puede ser *α*: hélice alfa, *β*: hoja beta o *γ*: lazo) en una proteína *p*, y |*ss(p)*| es la longitud de la estructura secundaria *ss* en la proteína *p*.

6.8 Análisis del efecto del sesgo del contenido de GC genómico en los parámetros conformacionales de los aminoácidos en la estructura secundaria de las proteínas

Para cada una de las 192 secuencias proteómicas consideradas en el primer grupo de estudio, se evaluaron los parámetros conformacionales de los aminoácidos para las estructuras secundarias usando el procedimiento original descrito por Chou y Fasman[2]. Es importante describir dos variables. La primera es la frecuencia de cada residuo de aminoácido presente en las estructuras secundarias (hélice alfas, hoja beta y lazos) en relación con la frecuencia de tales aminoácidos en el proteoma:

$$F_{ss}^{aa} = \frac{\sum_{p \in P(G)} |ss(p, aa)|}{\sum_{p \in P(G)} |p(aa)|},$$

donde *G* es el genoma de interés, *P(G)* es su proteoma, *aa* es un aminoácido dado, *p* es una proteína, |*ss(p,aa)*| es el número de veces que *aa* aparece en la estructura secundaria *ss* (puede ser *α*: alfa-hélice, *β*: hoja beta ó *γ*: lazo) en una proteína *p*, y |*p(aa)*| es el número de veces que *aa* aparece en la proteína *p*.

La segunda variable es la frecuencia relativa de los aminoácidos por cada estructura secundaria (hélice alfa, beta plagada y lazo) con relación al número de aminoácidos en el proteoma:

$$F_{ss} = \frac{\sum_{p \in P(G)} |ss(p)|}{|P(G)|},$$

donde *G* es el genoma de interés, *P(G)* es su proteoma, |*P(G)*| es su longitud, *p* es una proteína, y |*ss(p)*| es la longitud de la estructura secundaria *ss* (puede ser *α*: alfa-hélice, *β*: hoja beta ó *γ*: lazo) en la proteína *p*.

Los parámetros conformacionales (PC) fueron obtenidos cuando F_{ss}^{aa} es dividido por F_{ss} como se ve la siguiente fórmula:

$$PC = \frac{F_{ss}^{aa}}{F_{ss}}.$$

Los valores así obtenidos fueron graficados con respecto al contenido de GC de sus correspondientes secuencias genómicas.

6.9 Evaluación de las frecuencias de estructuras secundarias de los proteomas

Las frecuencias en las que los diferentes elementos de estructura secundaria, hélices alfa, hojas beta o lazos, representados en los proteomas fueron evaluados usando las siguientes fórmulas, respectivamente:

$$Alfa^{G} = \frac{\sum_{p \in P(G)} |\alpha(p)|}{|P(G)|},$$

$$Beta^G = \frac{\sum_{p \in P(G)} |\beta(p)|}{|P(G)|},$$

$$Coil^G = \frac{\sum_{p \in P(G)} |\gamma(p)|}{|P(G)|},$$

donde *G* es el genoma de interés, *P(G)* es su proteoma, |*P(G)*| es el número total de aminoácidos en el proteoma, *p* es una proteína dada del proteoma *P(G)*, *α(p)* es el número de aminoácidos encontrados en hélices alfa, *β(p)* es el número de aminoácidos en hojas beta y *γ(p)* es el número de aminoácidos que forman parte de los lazos.

6.10 Agrupación de proteínas en COGs

A partir de la base de datos KEGG GENOME[54], un conjunto de 1,544 procariotas representativo a nivel género fue elegido como el segundo grupo de estudio. Las proteínas de este conjunto de organismos fueron agrupadas en COGs (Clusters of Orthologous Genes)[56] mediante un análisis computacional de un algoritmo de cuatro pasos:

- Primero, para un COG dado en la base de datos COG de NCBI (https://www.ncbi.nlm.nih.gov/research/cog), se obtuvieron todas las proteínas correspondientes.

- Segundo, para cada conjunto de grupos COG, se alinearon las proteínas usando el programa MUSCLE[57].

- Tercero, para cada COG, un modelo oculto de Markov (HMM) fue construido usando el programa de *hmmerbuild* del paquete HMMER[58].

- Cuarto, usando las matrices generadas con HMM para cada COG, se evaluaron todas las secuencias de proteínas en nuestro grupo de organismos para identificar los dominios de proteínas que mejor corresponden al modelo.

6.11 Selección de secuencias de proteínas representativas por COGs

En el grupo de investigación del Dr. Merino se cuenta con una asignación de COGs hecha con modelos ocultos de Markov (HMM), descritos en el punto 6.10 de esta sección. Para seleccionar las secuencias de proteínas ortólogas que mejor representaran a cada COG, se obtuvo la distribución de las longitudes de las proteínas perteneciente a cada COG y fueron evaluadas la media y la desviación estándar. Los criterios son descritos a continuación:

- Como primer criterio para considerar a una proteína como representante de un COG para el análisis de estructura secundaria, solo se incluyó aquellas proteínas cuyas longitudes estuvieran ubicadas a no más de una desviación estándar a partir de la media de la distribución de longitud correspondiente.

- Para el segundo criterio de inclusión, se evaluó la distribución de las longitudes de los dominios COG identificados en las proteínas utilizando los correspondientes modelos ocultos de Markov (HMM) para cada COG. Aquí solo se incluyeron aquellas proteínas cuyas secuencias tuvieran una cobertura de al menos el 80% del valor promedio de la longitud del dominio COG.

La primera condición fue usada para excluir proteínas multidominio con regiones grandes de secuencia que no estuvieran asociadas con el COG a analizar y la segunda condición fue usada para descartar proteínas con dominios COG parciales.

6.12 Evaluación de las frecuencias de las estructuras secundarias de las proteínas COG

La estructura secundaria de las proteínas de los COG se predijo utilizando PSSPRED[55] versión 4. Con estos datos, se calculó la frecuencia de las estructuras secundarias: hélices alfa, hojas beta y lazos para cada uno de los 4,511 COGs de nuestro estudio, utilizando las siguientes fórmulas:

$$COG_{\alpha}^{p} = \frac{|\alpha(p)|}{|p|},$$

$$COG_{\beta}^{p} = \frac{|\beta(p)|}{|p|},$$

$$COG_{\gamma}^{p} = \frac{|\gamma(p)|}{|p|},$$

donde *p* es una proteína dada, |*p*| es la longitud de la secuencia de la proteína, y |*ss(p)*| es el número de aminoácidos predichos a estar presentes en una de las tres principales estructuras secundarias de proteínas (*α*: hélice alfa, *β*: hoja beta, *γ*: lazos).

6.13 Alineamientos múltiples de los elementos de estructura secundaria de proteínas ortólogas

Las estructuras secundarias de las proteínas se representaron como un código de tres letras, H, E y C para representar los elementos de hélice alfa, hoja beta y lazo, respectivamente. Los elementos de la estructura secundaria de cada COG fueron alineados usando el programa MUSCLE[57].

6.14 Lenguaje de programación usados

El pipeline generado para acceder y procesar los datos de la base de datos KEGG y los resultados obtenidos por el programa PSSPRED[55] fueron escritos en Perl, Python 3 y R y están disponibles en https://biocomputo.ibt.unam.mx/gcto2d/programs/. Para el análisis de los datos se usaron las paqueterías Pandas[59] y Numpy[60] de Python 3, para el análisis estadístico se usó *format.value* de la paquetería base de R.

En el laboratorio del Dr. Merino se desarrolló el servicio web GCto2D (https://biocomputo.ibt.unam.mx/gcto2d/) para poner a disposición de la comunidad científica nuestros resultados sobre el efecto del contenido de GC de los genes en las estructuras secundarias de las proteínas ortólogas. La página web GCto2D fue desarrollada usando HTML5/CSS como interfaz y se complementó con una combinación de vanilla JavaScript, PHP, Perl y MySQL para el backend y así garantizar una interacción rápida con los navegadores web modernos. El servicio de implementación está alojado en una instancia del servidor Apache HTTPD v2.4.

7. RESULTADOS

7.1 Diversidad del contenido de GC genómico a través de filos procariotas

A partir del primer grupo de estudio, el cual estaba conformado por 4,655 secuencias genómicas no redundantes de especies procariotas, se seleccionaron 192 basadas en su contenido de GC genómico, de las cuales 178 secuencias corresponden a bacterias y 14 a arqueas. El rango del contenido de GC genómico de estos organismos de estudio fue amplio, del 20 al 74% (Fig 1).

Los valores del contenido de GC genómico de nuestros 192 organismos de estudio fueron agrupados en 32 filos, de los cuales 28 eran de bacteria y solo 4 de

arqueas (Tabla S1). En la Fig 1 se pueden observar claras tendencias específicas para algunos filos:

- Actinobacteria fue el filo con contenido de GC genómico más alto, con una media de 70.42%, con miembros como *Cellulomonas fimi* ATCC 484 presentando el valor de contenido de GC de 74.6%. Acidobacteria también presentó un contenido de GC alto, con una media de 62.64% entre sus miembros.

- Fusobacteria y Tenericutes fueron los filos que presentaron valores con bajo contenido de GC genómico, con una media de 29.69% y 26.64%, respectivamente.

- El filo Proteobacteria merece una mención especial debido a que cuenta con el número más grande de organismos representativos secuenciados y con el rango del contenido de GC genómico más amplio. Se observó variación en la media del contenido de GC genómico entre las principales cuatro clases de Proteobacteria: Alpha, Beta, Delta y Gamma con valores de 54.41%, 58.67%, 57.37% y 45.53%, respectivamente.

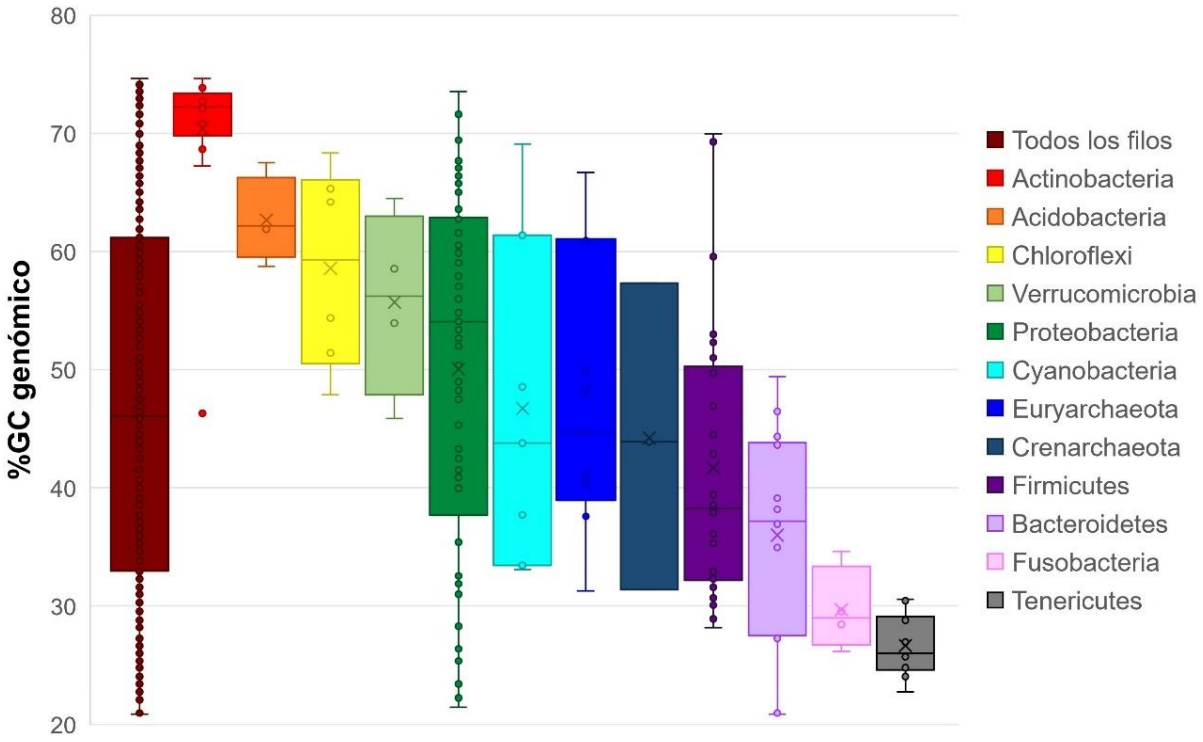


Figura 1. Distribución del contenido de GC genómico de 192 procariotas de estudio. Los boxplots muestran los resultados de 12 filos con al menos tres organismos con respecto a su contenido de GC genómico. El primer boxplot corresponde a los valores de contenido de GC genómico de todos los microorganismos de estudio, que van del 20 al 74%. Los boxplots fueron ordenados de mayor a menor de acuerdo con los valores de contenido de GC genómico del primer cuartil.

7.2 El contenido de GC genómico impone un sesgo en la frecuencia de aminoácidos del proteoma y las estructuras secundarias de sus proteínas

Debido a los diferentes contenidos de GC de los codones utilizados para codificar cada uno de los 20 aminoácidos, se ha documentado ampliamente que las frecuencias relativas de los aminoácidos de los proteomas varían a medida que fluctúa el contenido de GC genómico[1,16,19–22].

Diversos estudios han agrupado a los aminoácidos según el contenido de GC de los codones que los codifican y fueron analizadas sus frecuencias en los proteomas con respecto al GC genómico[16,19,20,22]. Es decir, con el incremento del contenido de GC genómico, la composición de los aminoácidos codificados por codones con alto

contenido de GC tendieron a incrementar (Ala, Gly y Pro), mientras que aquellos codificados por codones con bajo contenido de GC tendieron a disminuir (Lys, Ans e Ile).

En el presente estudio se categorizó a los aminoácidos en tres grupos: aquellos codificados por codones con contenido de GC alto, neutro y bajo (Tabla 1) y se realizó un estudio similar a los mencionados anteriormente utilizando un conjunto de 192 genomas procariotas, obteniendo resultados concordantes (Fig 2a, c, e y Fig S1). Para la mayoría de los aminoácidos, se observó una relación lineal entre el contenido de GC de sus codones y las frecuencias relativas con la cual son encontradas en los proteomas, mientras que algunos otros, tal como Gln y Met, los modelos de regresión polinomial más complejos se ajustan mejor a la curva (Fig S1 h.1, m.1, respectivamente).

En las gráficas de lado izquierdo de la Fig 2, la relación de la frecuencia de cada uno de los 20 aminoácidos, con respecto al contenido de GC genómico de 192 procariotas, mostró lo siguiente:

- Una regresión lineal con pendiente positiva (Fig 2a y Fig S1 a.1 - d.1, Tabla S2) para los aminoácidos codificados por codones ricos en contenido de GC (Ala, Gly, Pro, Arg y Trp; Tabla 1).

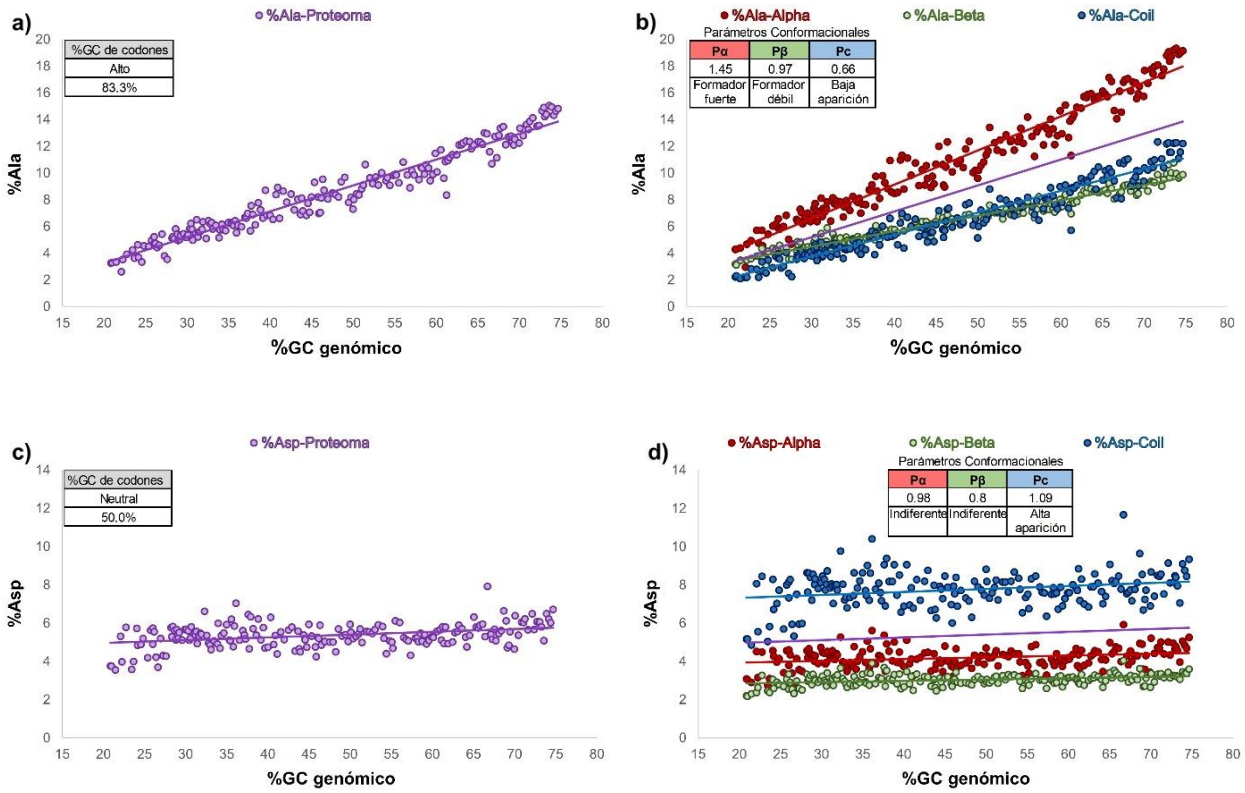
- Los aminoácidos codificados por codones neutros en contenido de GC (His, Asp, Thr, Cys, Glu y Ser; Tabla 1) parecen no estar afectados de manera importante por el contenido de GC genómico (Fig 2c y Fig S1 f.1 – g.1, i.1 – k.1, Tabla S2).

- Una regresión lineal con pendiente negativa (Fig 2e y S1 n.1 – q.1, Tabla S2) para los aminoácidos codificados por codones bajo en contenido de GC (Phe, Tyr, Lys, Asn e Ile; Tabla 1).

Cabe señalar que Val y Leu presentan un comportamiento inusual en sus pendientes con respecto al grupo de aminoácidos al que pertenecen. Esto es, Val (aminoácido con codones neutros en contenido de GC) presenta una correlación positiva

con respecto al incremento del contenido de GC genómico (Fig S1 e.1 y Tabla S2), mientras que Leu (aminoácido con codones bajos en contenido de GC) también presenta tendencia positiva (Fig S1 l.1 y Tabla S2).

Otro punto importante por mencionar es que debido a las propiedades fisicoquímicas de los aminoácidos, estos presentan tendencias específicas que previenen (interruptores), contribuyen (formadores) o tienen un efecto neutral (indiferentes) en la formación de las estructuras secundarias de las proteínas[2,3,24,45,61]. En este sentido, repetimos el análisis descrito anteriormente para las secuencias del proteoma, pero consideramos de forma independiente los aminoácidos presentes en los elementos de estructuras secundarias. Este análisis se muestra como gráficas de la Fig 2b, d, f y Fig S1 a.2 – q.2, además se reportaron los datos de regresión lineal como Tabla S3.



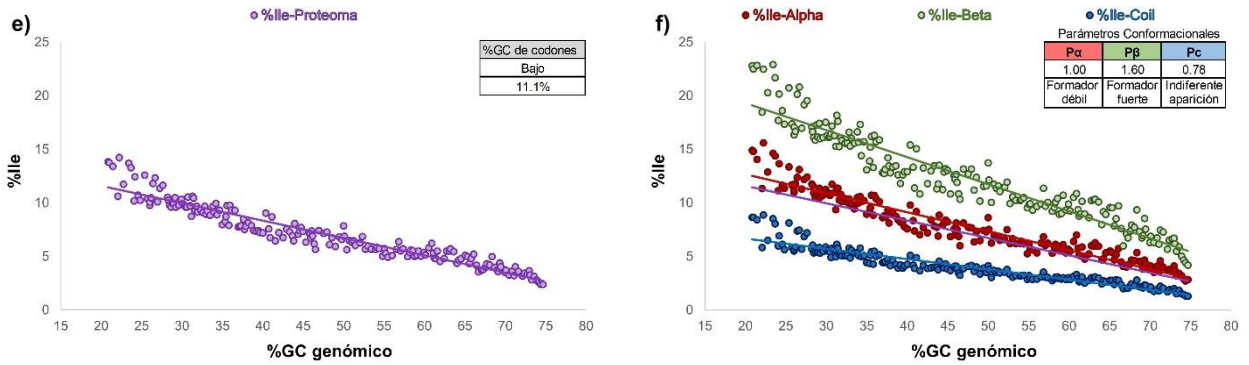


Figura 2. Efecto del sesgo del contenido de GC genómico sobre las frecuencias de los aminoácidos en el proteoma y las estructuras secundarias de proteínas. Las gráficas representan a las frecuencias relativas de los aminoácidos en los proteomas (a, c, e) o en las estructuras secundarias de las proteínas (b, d, f) de 192 procariotas. Los aminoácidos de las figuras fueron elegidos considerando sus tendencias a ser parte de cada una de las estructuras secundarias de las proteínas (b, formador fuerte en hélices alfa; d, indiferente en las tres estructuras secundarias; f, formador fuerte en hojas beta) o considerando el contenido de GC de los codones por los que están codificados (a, c, e, aminoácidos con codones alto, neutro y bajo contenido de GC, respectivamente). Como referencia, las líneas de regresión de la frecuencia de aminoácidos obtenidas en el análisis del proteoma son mostradas en morado (b, d, f). Los parámetros conformacionales de los aminoácidos descritos por Chou y Fasman[2,3] son mostrados en la parte superior de cada gráfica.

En la Fig 2b, d, f se muestran tres ejemplos de la relación entre la frecuencia de un aminoácido en una estructura secundaria en específico con respecto al contenido de GC genómico de nuestros 192 procariotas de estudio y observamos tres principales características:

- Los aminoácidos que tienden a contribuir a la formación (formadores) de una estructura secundaria específica de las proteínas presentaron valores de pendiente más altos que la pendiente de la línea de regresión del proteoma.

- Por otro lado, los aminoácidos que tienden a prevenir (interruptores) la formación de estructuras secundarias de las proteínas presentan valores de pendiente más bajos con respecto a la pendiente de la línea de regresión del proteoma.

- Adicionalmente, aquellos aminoácidos con un efecto neutral (indiferentes) en la formación de las estructuras secundarias presentaron valores de pendientes de línea de regresión similares a los obtenidos en el análisis del proteoma completo.

Un ejemplo representativo de la explicación anterior es el caso del aminoácido Ala, donde la pendiente más alta de la línea de regresión se presenta en la estructura secundaria hélice alfa (línea roja, Fig 2b), quedando por arriba de la pendiente de línea de regresión obtenida por el proteoma (línea morada, Fig 2b). Esto es indicativo de la preferencia de este aminoácido por encontrarse en este tipo de estructura secundaria, mientras que las pendientes más pequeñas para hoja beta y lazo (líneas verde y azul, respectivamente, Fig 2b) indican la baja propensión de Ala a formar parte de las estructuras secundarias antes mencionadas.

7.3 Análisis del efecto del sesgo del contenido de GC genómico en los parámetros conformacionales de los aminoácidos en la estructura secundaria de proteínas

Las tendencias de los aminoácidos a formar parte de una estructura secundaria específica han sido caracterizadas desde 1974 por Chou y Fasman y se expresan como parámetros conformacionales (PC), estos son valores asignados que van de 0 a 2[2,3]. En tales estudios, los PC fueron evaluados considerando dos variables:

- Las frecuencias de cada uno de los aminoácidos presentes en las hélices alfa, hojas beta y lazos en relación con las frecuencias de dichos aminoácidos en el proteoma.

- Las frecuencias relativas de los aminoácidos para cada estructura secundaria en relación con el número de aminoácidos en el proteoma.

Cabe destacar que el número de proteínas utilizadas en el estudio pionero de Chou y Fasman[2] fue de solo 15. A diferencia del estudio mencionado, en el presente estudio, actualizamos estos valores conformacionales considerando las secuencias de aminoácidos de 192 proteomas completos que se seleccionaron en función del contenido de GC de sus secuencias genómicas correspondientes. Inesperadamente, observamos que los valores conformacionales de algunos aminoácidos no eran constantes. Es decir,

nuestros PC calculados presentaron variaciones pequeñas, pero estadísticamente significativas en función de los valores de contenido de GC genómico (Fig 3).

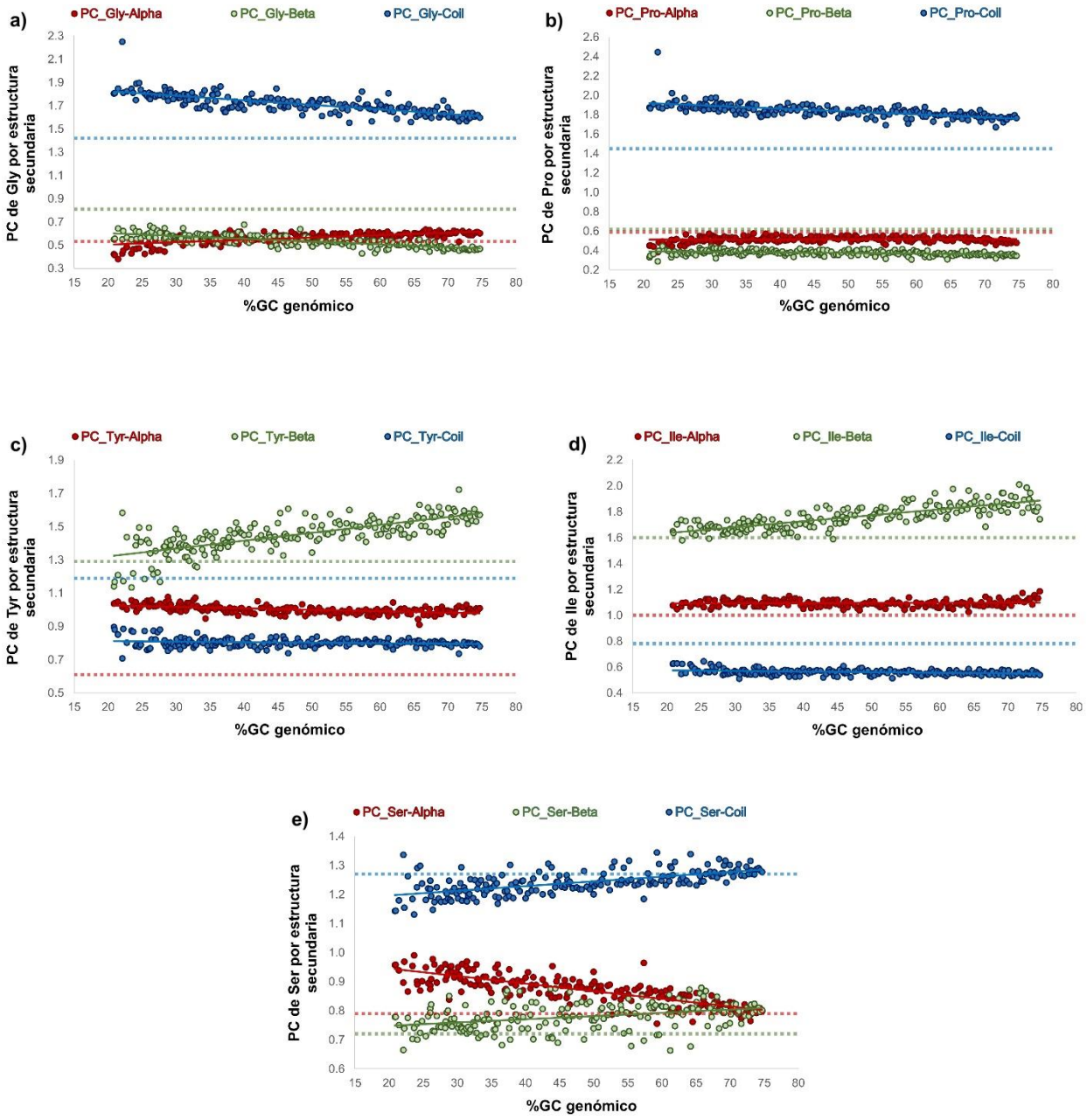


Figura 3. Efecto del contenido de GC genómico sobre los valores de los parámetros conformacionales de aminoácidos. Los valores de PC para Gly, Pro, Tyr, Ile y Ser en hélices alfa (círculos rojos), hoja beta (círculos verdes) y lazos (círculos azules) fueron evaluados en función del contenido de GC genómico. Las líneas de regresión de las estructuras secundarias están representadas por líneas continuas, mientras que los valores de PC reportados por Chou y Fasman[2,3] están representados por líneas punteadas.

De acuerdo con los resultados obtenidos en nuestro estudio, los aminoácidos cuyos valores de PC que presentan mayor diferencia significativa a medida que varía el contenido de GC genómico son: Gly y Pro encontrados en lazos (Fig 3a-b, respectivamente); Tyr, Ile, Ala, Asn y Leu en hojas beta (Fig 3c-d y Fig S2 a-c, respectivamente); y Ser en hélices alfa (Fig 3e). El resto de las figuras de aminoácidos se muestran en la Fig S2, y los datos de regresión lineal de todos los aminoácidos se presentan en la Tabla S4.

7.4 El contenido de GC genómico de organismos procariotas impone un sesgo en las frecuencias de estructuras secundarias de los proteomas

Considerando que el contenido de GC genómico en 192 procariotas: (a) puede variar de manera importante de acuerdo con la filogenia de los organismos (Fig 1), (b) afecta las frecuencias totales de los aminoácidos de los proteomas y en sus correspondientes estructuras secundarias (Fig 2), y (c) impone un sesgo en los parámetros conformacionales de aminoácidos (Fig 3). Aquí analizamos si las frecuencias totales de cada estructura secundaria de las proteínas en los proteomas varían en función del contenido de GC de los genomas por los que están codificados (Fig 4).

Nuestros resultados indican que la composición de estructuras secundarias de los proteomas no es universal, sino que presenta variaciones que se correlacionan con el contenido de GC genómico. A medida que aumenta el contenido de GC de los genomas, las frecuencias relativas de los lazos tienden a aumentar, mientras que las frecuencias relativas de las hélices alfa y las hojas beta tienden a disminuir (Fig 4, los datos de regresión lineal en Tabla S5).

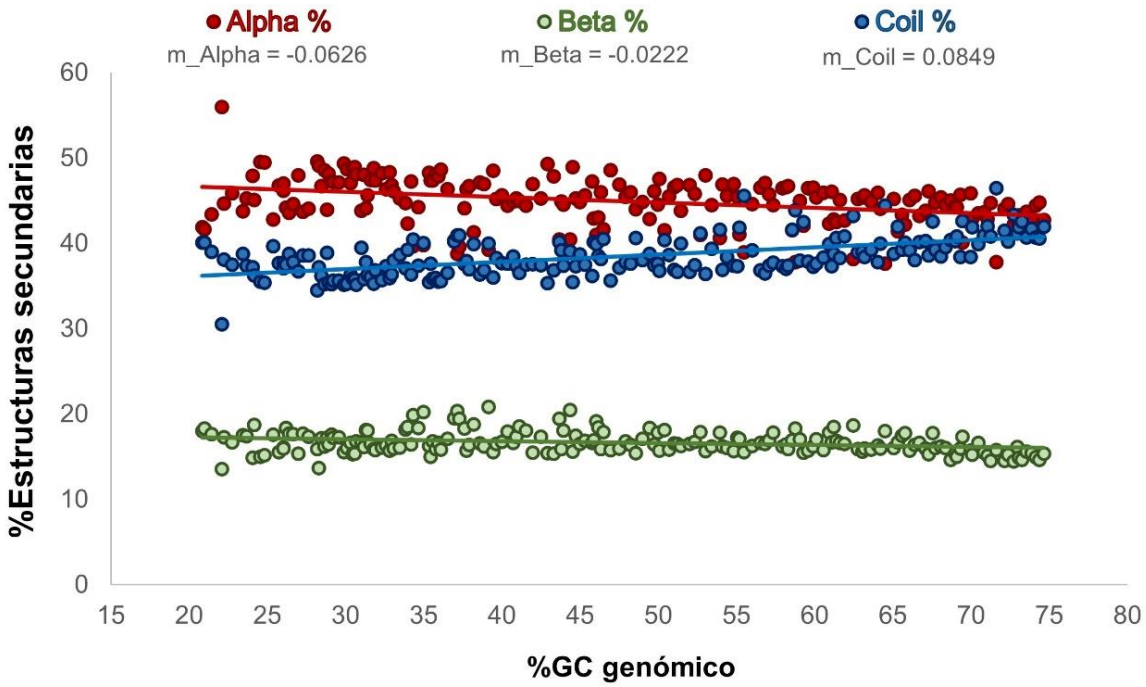


Figura 4. Frecuencia relativa de las estructuras secundarias de los proteomas en función del contenido de GC genómico. Las frecuencias relativas de las estructuras secundarias: hélice alfa, hoja beta y lazo en nuestros organismos de estudio fueron evaluadas y graficadas con respecto a su correspondiente contenido de GC genómico. Se obtuvo una regresión lineal para cada uno de los elementos de estructura secundaria. El valor de la pendiente (m) para cada regresión lineal se indica en parte superior de la figura.

7.5 El contenido de GC de los genes impone un sesgo en las estructuras secundarias de algunas familias de proteínas ortólogas (COGs)

Los cambios pequeños, pero estadísticamente significativos en la frecuencia relativa de las estructuras secundarias de los proteomas en función del contenido de GC genómico, identificado y mostrado en la Fig 4, pueden explicarse por al menos tres razones diferentes:

a. Por la presencia de diferentes proteínas en los proteomas;

b. Por un cambio en las frecuencias relativas de las estructuras secundarias en proteínas ortólogas; o

c. Por una combinación de las posibilidades anteriores.

Con el fin de determinar cuál de estas posibilidades (a., b., ó c.) es la más acertada, repetimos nuestro análisis del efecto de sesgo del contenido de GC en la estructura secundaria de las proteínas considerando diferentes conjuntos de genes ortólogos. Para tener más confianza en nuestros análisis, ampliamos el conjunto inicial de organismos de referencia para incluir 1,544 procariotas representativas a nivel de género.

En este nuevo análisis, las proteínas ortólogas se agruparon teniendo en cuenta la clasificación de la base de datos COG[56]. Para garantizar que las comparaciones entre proteínas se restringieran principalmente a dominios homólogos, utilizamos criterios de inclusión estrictos para la selección de proteínas en los grupos COG (Materiales y Métodos).

Las gráficas del contenido de GC de genes ortólogos contra las frecuencias relativas de sus estructuras secundarias revelaron que, para la mayoría de los COGs, no

hubo una variación significativa en las estructuras secundarias, tal como se muestra en el COG0002 (Fig. 5a, los datos de regresión lineal en la Tabla S6).

En contraste con el resultado anterior, nuestro estudio también mostró que casi el 5% de las secuencias COG analizadas presentaban variaciones pequeñas, pero estadísticamente significativas, en las estructuras secundarias de sus proteínas a medida que variaba el contenido de GC de sus genes correspondientes. Un ejemplo de este tipo se presenta en el COG3228 (Fig 5b y los datos de regresión lineal en la Tabla S6).

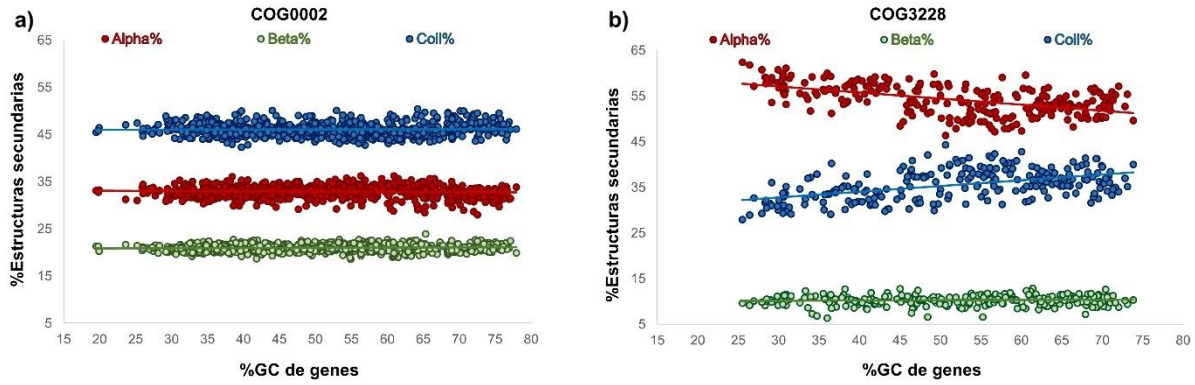


Figura 5. Regresión lineal comparativa entre COGs, sin y con sesgo impuesto por el contenido de GC de sus respectivos genes. a) COG0002: Acetilglutamato semialdehído deshidrogenasa y b) COG3228: Proteína no caracterizada conservada en bacterias. Las líneas de regresión para cada estructura secundaria se muestran como líneas sólidas.

Además de los resultados presentados en la Fig 5, representamos las hélices alfa, las hojas beta y los lazos con las letras H, E y C, respectivamente, y realizamos alineamientos múltiples de las estructuras secundarias de las proteínas ortólogas. Esto se hizo con el motivo de observar qué parte de la secuencia de proteínas ortólogas es mayormente sesgado por el contenido de GC de los genes que las codifican.

Las secuencias alineadas muestran una conservación significativa de la estructura secundaria en la mayoría de los grupos de proteínas ortólogas (COGs), este efecto es independiente de las importantes diferencias en los contenidos de GC de sus

respectivos genes que codifican proteínas, como se observa en los alineamientos múltiples de las estructuras secundarias del COG0002 (Fig S3).

Por otro lado, también mostramos los alineamientos múltiples de secuencias del COG3228 como un ejemplo de proteínas ortólogas que presentó variaciones en sus estructuras secundarias a medida que variaba el contenido de GC de sus genes (Fig S4). Para este tipo de COGs, observamos que los extremos amino- y carboxi- terminal de sus proteínas son las regiones más propensas a variaciones en la longitud y composición de sus estructuras secundarias.

El conjunto completo de gráficos para todos los COGs analizados y sus alineamientos múltiples de estructura secundaria de proteínas, están disponibles en nuestro servidor web GCto2D (https://biocomputo.ibt.unam.mx/gcto2d/).

8. CONCLUSIONES

Una de las principales características de los genomas de los organismos procariotas es la gran variación en la frecuencia con la que se utilizan las bases Guanina-Citosina (GC) en sus secuencias de DNA genómico. Estudios pioneros han demostrado el impacto del contenido de GC genómico en la distribución filogenética y en la frecuencia relativa de aminoácidos en los proteomas de los organismos[1,16,19–22].

Como primer punto, nuestro trabajo enriqueció estudios sobre la variación del contenido de GC genómico a través de diferentes filos procariotas, usando bases de datos actualizadas. Se observó que algunos filos no solo comparten un origen evolutivo, sino que están compuestos por microorganismos con características en común como el contenido de GC genómico, como es el caso de las Actinobacterias y Tenericutes.

Después de varias décadas de los primeros informes, el presente estudio da un paso más y analiza el efecto del contenido genómico de GC en la estructura secundaria de las proteínas. Como segundo punto demostramos, a través de un estudio

bioinformático, que la tendencia de un aminoácido a formar parte de una estructura secundaria varía en función del contenido de GC genómico. Además, como tercer punto, nuestro estudio muestra que los parámetros conformacionales de los aminoácidos en las estructuras secundarias de las proteínas: hélices alfa, hojas beta y lazos, no son constantes como se esperaba en un principio, sino que presentan variaciones estadísticamente significativas según el contenido de GC genómico.

Dado lo encontrado en el párrafo anterior, podemos decir que los parámetros conformacionales de Chou y Fasman han sido la referencia fundamental en cientos de estudios para predecir las estructuras secundarias de las proteínas. Sin embargo, estos parámetros conformacionales pueden ser visto faltos de rigurosidad al no tomar en cuenta otras características importantes como la composición total de proteínas (proteoma) y la influencia de su contenido de GC genómico de un organismo.

Adicionalmente, como punto cuatro, encontramos que la composición de las estructuras secundarias de los proteomas varía en relación con el contenido de GC genómico: los lazos aumentan a medida que aumenta el contenido de GC genómico, mientras que las hélices alfa y las hojas beta presentan una relación inversa. Con respecto a que los lazos de las proteínas aumentan al incrementar el contenido de GC genómico, se espera que las proteínas serán más propensas a tener regiones o proteínas intrínsecamente desordenadas (IDR o IDP, respectivamente).

Finalmente, como quinto punto, descubrimos que para la mayoría de los grupos de proteínas ortólogas la composición de estructuras secundarias parece mayormente invariante a pesar de que el contenido de GC de los genes que las codifican presenten variaciones. No obstante, identificamos que para algunos grupos particulares de proteínas ortólogas, el contenido de GC de los genes impone un sesgo en la composición de las estructuras secundarias de las proteínas que codifican.

9. PERSPECTIVAS

Recapitulando, el contenido de GC genómico es una característica biológica importante en todos los organismos, capaz de influenciar en rasgos esenciales como lo es el tamaño del genoma, los elementos genéticos (plásmidos), la relación filogenética, la adaptación a un ambiente, la composición de la estructura primaria y secundaria de las proteínas (nuestro proyecto), entre otras. De aquí se deriva la relevancia de seguir incrementando el conocimiento del efecto del contenido de GC genómico en otras propuestas abordando diferentes niveles estructurales de las proteínas.

Nuestro estudio del efecto del contenido de GC genómico puede ser complementado con un árbol filogenético que muestre los ambientes de los cuales provienen los organismos de estudio, el estilo de vida al que pertenecen, el tamaño del genoma y otras características distintivas. Además, ya que se cuenta con la predicción de estructuras secundarias, el contenido de GC genómico también puede ser extendido analizando el uso codónico de los aminoácidos de los procariotas de estudio.

Apoyándonos de la premisa de que el GC genómico impacta en una minoría de grupos de proteínas ortólogas, nuestro trabajo puede ampliarse, de manera que se estudie el contenido de GC y el posible impacto en los elementos del pangenoma: genes núcleo, genes dispensables y genes únicos[62].

Otra manera que puede ser abordado el estudio del GC genómico es utilizando herramientas de predicción estructural de proteínas (como AlphaFold[63]). De esta manera se visualizaría mejor qué partes de las proteínas, hablando en términos de estructura primaria y secundaria, se ve principalmente afectado por el contenido de GC genómico.

Finalmente, nuestro trabajo al estudiar el sesgo del GC genómico en las estructuras secundarias de las proteínas, abre la posibilidad de abarcar nuevas preguntas de investigación en los niveles superiores de organización de las proteínas:

estructura terciaria (influencia en zonas internas y externas) y cuaternaria (influencia en monómeros y zonas de contacto entre monómeros).

10. APÉNDICE

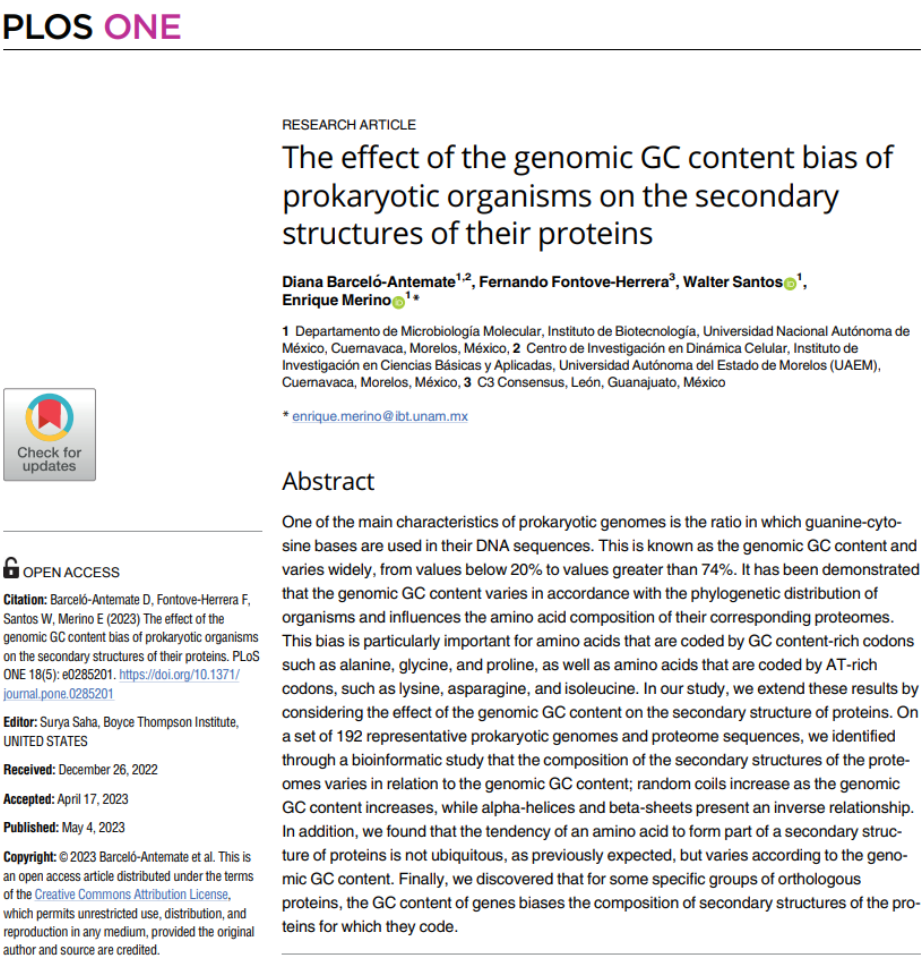
El artículo científico fue enviado, aceptado y publicado en la revista PLOS ONE.

Se adjunta el link del artículo:
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0285201

Se adjunta la cita del artículo:

The effect of the genomic GC content bias of prokaryotic organisms on the secondary structures of their proteins. Barceló-Antemate D, Fontove-Herrera F, Santos W, Merino E (2023). PLOS ONE 18(5): e0285201. https://doi.org/10.1371/journal.pone.0285201

También se adjunta la portada el artículo:



PLOS ONE

RESEARCH ARTICLE

The effect of the genomic GC content bias of prokaryotic organisms on the secondary structures of their proteins

Diana Barceló-Antemate[1,2], Fernando Fontove-Herrera[3], Walter Santos[1], Enrique Merino[1]*

1 Departamento de Microbiología Molecular, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, México, 2 Centro de Investigación en Dinámica Celular, Instituto de Investigación en Ciencias Básicas y Aplicadas, Universidad Autónoma del Estado de Morelos (UAEM), Cuernavaca, Morelos, México, 3 C3 Consensus, León, Guanajuato, México

* enrique.merino@ibt.unam.mx

Check for updates

OPEN ACCESS

Citation: Barceló-Antemate D, Fontove-Herrera F, Santos W, Merino E (2023) The effect of the genomic GC content bias of prokaryotic organisms on the secondary structures of their proteins. PLoS ONE 18(5): e0285201. https://doi.org/10.1371/journal.pone.0285201

Editor: Surya Saha, Boyce Thompson Institute, UNITED STATES

Received: December 26, 2022

Accepted: April 17, 2023

Published: May 4, 2023

Copyright: © 2023 Barceló-Antemate et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

One of the main characteristics of prokaryotic genomes is the ratio in which guanine-cytosine bases are used in their DNA sequences. This is known as the genomic GC content and varies widely, from values below 20% to values greater than 74%. It has been demonstrated that the genomic GC content varies in accordance with the phylogenetic distribution of organisms and influences the amino acid composition of their corresponding proteomes. This bias is particularly important for amino acids that are coded by GC content-rich codons such as alanine, glycine, and proline, as well as amino acids that are coded by AT-rich codons, such as lysine, asparagine, and isoleucine. In our study, we extend these results by considering the effect of the genomic GC content on the secondary structure of proteins. On a set of 192 representative prokaryotic genomes and proteome sequences, we identified through a bioinformatic study that the composition of the secondary structures of the proteomes varies in relation to the genomic GC content; random coils increase as the genomic GC content increases, while alpha-helices and beta-sheets present an inverse relationship. In addition, we found that the tendency of an amino acid to form part of a secondary structure of proteins is not ubiquitous, as previously expected, but varies according to the genomic GC content. Finally, we discovered that for some specific groups of orthologous proteins, the GC content of genes biases the composition of secondary structures of the proteins for which they code.

11. MATERIAL COMPLEMENTARIO

Tablas Complementarias

Tabla S1. Clasificación taxonómica de 192 procariotas con su contenido de GC genómico. El dominio, filo y especie es mostrado en la primera, segunda y tercera columna, respectivamente. El ID KEGG es encontrado en la cuarta columna. Los organismos de estudio están en orden ascendente con respecto al contenido de GC genómico (quinta columna).

Dominio	Filo	Especie	ID_KEGG	%GC genómico
Bacteria	Bacteroidetes	Candidatus Sulcia muelleri PSPU	smup	20.8321
Bacteria	Bacteroidetes	Candidatus Sulcia muelleri CARI	sum	20.9465
Bacteria	Proteobacteria	Buchnera aphidicola BCc, endosymbiont of Cinara cedri	bcc [g]	21.4278
Bacteria	Unclasiffied	Bacterium AB1	baab	22.0578
Bacteria	Proteobacteria	Candidatus Purcelliella pentastirinorum OLIH	ppet [g]	22.2156
Bacteria	Tenericutes	Candidatus Hepatoplasma crinochetorum Av	hcr	22.7287
Bacteria	Proteobacteria	Candidatus Portiera aleyrodidarum AF-CAI	pli [g]	23.4017
Bacteria	Proteobacteria	Wigglesworthia glossinidia (Wigglesworthia brevipalpis), endosymbiont of Glossina brevipalpis	wbr [g]	23.6454
Bacteria	Tenericutes	Mycoplasma capricolum subsp. capripneumoniae 9231-Abomsa	mcac	24.0394
Archaea	Nanoarchaeota	Candidatus Nanopusillus acidilobi	naa	24.1400
Bacteria	Tenericutes	Candidatus Phytoplasma ziziphi Jwb-nky	pzi	24.5311
Bacteria	Tenericutes	Spiroplasma floricola 23-6	sfz	24.7916
Bacteria	Proteobacteria	Candidatus Profftella armatura	ssdc [b]	25.3511
Bacteria	Tenericutes	Ureaplasma parvum serovar 3 ATCC 700970	uur	25.7104
Bacteria	Tenericutes	Ureaplasma urealyticum serovar 10 ATCC 33699	uue	25.9949
Bacteria	Tenericutes	Spiroplasma kunkelii CR2-3x	skn	25.9949
Bacteria	Fusobacteria	Streptobacillus moniliformis DSM 12112	smf	26.1605
Bacteria	Proteobacteria	Buchnera aphidicola USDA, endosymbion of Myzus persicae	bapu [g]	26.3736
Bacteria	Proteobacteria	Wigglesworthia glossinidia endosymbiont of Glossina morsitans morsitans	wgl [g]	26.6267
Bacteria	Tenericutes	Mesoplasma tabanidae BARC 857	mtab	26.9512
Bacteria	Bacteroidetes	Blattabacterium sp. Bge, endosymbiont of Blattella germanica	bbl	27.2471
Bacteria	Bacteroidetes	Blattabacterium sp. (Mastotermes darwiniensis) MADAR, endosymbiont of Mastotermes darwiniensis	bmm	27.5994
Bacteria	Firmicutes	Paeniclostridium sordellii AM370	psor	28.1833
Bacteria	Proteobacteria	Candidatus Babela massiliensis BABL1 (delta proteobacterium BABL1)	dpb [d]	28.2811
Bacteria	Fusobacteria	Sneathia amnii Sn35	sns	28.4298
Bacteria	Proteobacteria	Arcobacter nitrofigilis DSM 7299	ant	28.6183
Bacteria	Tenericutes	Candidatus Mycoplasma girerdii VCU_M1	mgj	28.7858
Bacteria	Firmicutes	Parvimonas micra KCOM 1535	pmic	28.9066
Bacteria	Thermotogae	Marinitoga piezophila KA3	mpz	29.1024
Bacteria	Tenericutes	Acholeplasma palmae J233	apal	29.1959
Bacteria	Fusobacteria	Fusobacterium mortiferum ATCC 9817	fmo	29.5523
Bacteria	Spirochaetes	Borrelia turicatae 91E135	btu	29.8880
Bacteria	Firmicutes	Clostridioides difficile 630 (Clostridium difficile 630)	cdf	30.0737
Bacteria	Firmicutes	Clostridium saccharoperbutylacetonicum N1-4(HMT)	csr	30.2976
Bacteria	Tenericutes	Mesoplasma syrphidae YJS	msyr	30.4348
Bacteria	Tenericutes	Spiroplasma eriocheiris DSM 21848	seri	30.5609
Bacteria	Firmicutes	Gottschalkia acidurici 9a (Clostridium acidurici 9a)	cad	30.6765
Bacteria	Proteobacteria	Ehrlichia canis Jake	ecn [a]	30.9901
Bacteria	Proteobacteria	Campylobacter coli OR12	ccoo	31.1465
Archaea	Euryarchaeota	Methanothermococcus okinawensis IH1	mok	31.2872
Archaea	Crenarchaeota	Acidianus manzaensis YN-25	aman	31.3988
Bacteria	Firmicutes	Gemella sp. oral taxon 928	got	31.5735
Bacteria	Firmicutes	Clostridium ljungdahlii DSM 13528	clj	31.7764
Bacteria	Proteobacteria	Francisella halioticida DSM 23729	fha [g]	31.8845
Bacteria	Firmicutes	Finegoldia magna ATCC 29328	fma	32.2942
Bacteria	Proteobacteria	Allofrancisella guangzhouensis (Francisella guangzhouensis 08HL01032)	fgu [g]	32.5550
Bacteria	Proteobacteria	Rickettsia sp. MEAM1 (Bemisia tabaci)	ric [a]	32.7790
Bacteria	Firmicutes	Melissococcus plutonius DAT561	mpx	32.9302
Bacteria	Cyanobacteria	Candidatus Atelocyanobacterium thalassa (Cyanobacterium UCYN-A)	cyu	33.0668
Bacteria	Cyanobacteria	Geminocystis sp. NIES-3708	gee	33.4436
Bacteria	Dictyoglomi	Dictyoglomus thermophilum H-6-12	dth	33.8117
Archaea	Thaumarchaeota	Candidatus Nitrosopumilus adriaticus NF5	nin	33.9370
Bacteria	Nitrospirae	Thermodesulfovibrio yellowstonii DSM 11347	tye	34.1596

Tabla S1. Continuación…

Dominio	Filo	Especie	ID_KEGG	%GC genómico
Bacteria	Ignavibacteriae	Ignavibacterium album JCM 16511	ial	34.3083
Bacteria	Fusobacteria	Sebaldella termitidis ATCC 33386	str	34.6157
Bacteria	Bacteroidetes	Winogradskyella sp. J14-2	wij	34.9474
Bacteria	Firmicutes	Turicibacter sp. H121	tur	35.3083
Bacteria	Proteobacteria	Wolbachia sp. wRi, endosymbiont of Drosophila simulans	wri [a]	35.4029
Bacteria	Firmicutes	Thermoanaerobacterium saccharolyticum JW/SL-YS485	tsh	35.5487
Bacteria	Firmicutes	Bacillus thuringiensis HD-789	btn	35.7603
Bacteria	Deferribacteres	Calditerrivibrio nitroreducens DSM 19672	cni	35.9156
Bacteria	Firmicutes	Anaerococcus prevotii DSM 20548	apr	36.0897
Bacteria	Firmicutes	Caldicellulosiruptor hydrothermalis 108	chd	36.5166
Bacteria	Bacteroidetes	Formosa sp. Hel3_A1_48	foh	36.9274
Bacteria	Bacteroidetes	Pseudopedobacter saltans DSM 12145 (Pedobacter saltans DSM 12145)	psn	37.1201
Bacteria	Bacteroidetes	Arachidicoccus sp. KIS59-12	ark	37.2681
Archaea	Euryarchaeota	Methanohalobium evestigatum Z-7303	mev	37.5836
Bacteria	Cyanobacteria	Stanieria sp. NIES-3757	stan	37.7091
Bacteria	Firmicutes	Lactobacillus heilongjiangensis DSM 28069	lhi	37.8905
Bacteria	Bacteroidetes	Fermentimonas caenicola	pbt	38.1817
Bacteria	Firmicutes	Listeria monocytogenes 07PF0776 (serotype 4b)	lmp	38.5580
Bacteria	Firmicutes	Lactobacillus amylovorus GRL1118	lay	38.8181
Bacteria	Bacteroidetes	Capnocytophaga sp. ChDC OS43	capn	39.1228
Bacteria	Firmicutes	Leuconostoc citreum KM20	lci	39.4209
Bacteria	Chlamydiae	Chlamydia psittaci 84/55	cpsb	39.6117
Bacteria	Proteobacteria	Helicobacter pylori 2018	hpw	39.9743
Archaea	Euryarchaeota	Methanosalsum zhilinae DSM 4017	mzh	40.3322
Bacteria	Elusimicrobia	Elusimicrobium minutum Pei191	emi	40.6903
Bacteria	Proteobacteria	Pseudoalteromonas translucida KMM 520	ptn [g]	40.8837
Archaea	Euryarchaeota	Pyrococcus furiosus COM1	pfi	41.1008
Bacteria	Proteobacteria	Pseudoalteromonas spongiae UST010723-006	pspo [g]	41.5243
Bacteria	Proteobacteria	Candidatus Paracaedibacter acanthamoebae PRA3	paca [a]	41.9552
Bacteria	Proteobacteria	Cycloclasticus sp. P1	cyq [g]	42.4815
Bacteria	Firmicutes	Lentibacillus amyloliquefaciens LAM0015	lao	42.8933
Bacteria	Proteobacteria	Coxiella burnetii RSA 331	cbs [g]	43.2978
Bacteria	Bacteroidetes	Bacteroides caecimuris I48	bcae	43.6398
Bacteria	Cyanobacteria	Nostocales cyanobacterium HT-58-2	ncn	43.7823
Archaea	Crenarchaeota	Caldivirga maquilingensis IC-167	cma	43.9167
Bacteria	Bacteroidetes	Odoribacter splanchnicus DSM 20712	osp	44.3467
Bacteria	Firmicutes	Bacillus subtilis subsp. spizizenii TU-B-10	bst	44.5173
Archaea	Euryarchaeota	Ferroglobus placidus DSM 10642	fpl	44.7096
Bacteria	Candidatus Saccharibacteria	Candidatus Saccharibacteria oral taxon TM7x	sox	44.9521
Bacteria	Proteobacteria	Shewanella putrefaciens CN-32	spc [g]	45.3007
Bacteria	Synergistetes	Aminobacterium colombiense DSM 12261	aco	45.7533
Bacteria	Verrucomicrobia	Methylacidiphilum infernorum V4	min	45.8507
Bacteria	Calditrichaeota	Calditrix abyssi DSM 13497	caby	46.0097
Bacteria	Chlorobi	Chloroherpeton thalassium ATCC 35110	cts	46.1385
Bacteria	Actinobacteria	Atopobium parvulum DSM 20469	apv	46.3081
Bacteria	Bacteroidetes	Saprospira grandis Lewin	sgn	46.4727
Bacteria	Firmicutes	Bacillus infantis NRRL B-14911	bif	46.9142
Bacteria	Proteobacteria	Desulfotalea psychrophila LSv54	dps [d]	47.4956
Bacteria	Chloroflexi	Dehalococcoides mccartyi DCMB5	dmd	47.8840
Bacteria	Proteobacteria	Helicobacter heilmannii ASB1.4	hhm	48.2185
Bacteria	Cyanobacteria	Pseudanabaena sp. PCC 7367	pseu	48.5366
Bacteria	Proteobacteria	Shewanella sp. MR-4	she [g]	48.9649
Bacteria	Bacteroidetes	Porphyromonas gingivalis ATCC 33277	pgn	49.4097
Bacteria	Firmicutes	Paenibacillus swuensis DY6	pswu	49.7223
Archaea	Euryarchaeota	Methanothermobacter wolfeii SIV6	mwo	49.9596
Bacteria	Firmicutes	Clostridium bolteae ATCC BAA-613	cbol	50.0545
Bacteria	Proteobacteria	Bdellovibrio bacteriovorus Tiberius	bbat	50.3736
Bacteria	Proteobacteria	Desulfomonile tiedjei DSM 6799	dti [d]	50.6788
Bacteria	Firmicutes	Acidaminococcus intestini RyC-MR95	ain	51.0016
Bacteria	Firmicutes	Selenomonas ruminantium subsp. lactilytica TAM6421	sri	51.2020
Bacteria	Chloroflexi	Herpetosiphon aurantiacus DSM 785	hau	51.4149
Bacteria	Proteobacteria	Mariprofundus aestuarium CP-5	maes	51.9917
Bacteria	Firmicutes	Christensenella minuta DSM 22607	cmiu	52.3041

Tabla S1. Continuación…

Dominio	Filo	Especie	ID_KEGG	%GC genómico
Bacteria	Proteobacteria	Cellvibrio japonicus Ueda107	cja [g]	52.6590
Bacteria	Firmicutes	Lactobacillus fermentum F-6	lff	52.9948
Bacteria	Proteobacteria	Granulosicoccus antarcticus IMCC3135	gai [g]	53.3986
Bacteria	Verrucomicrobia	Coraliomargarita akajimensis DSM 45221	caa	53.9251
Bacteria	Proteobacteria	Citrobacter sp. CFNIH10	cir [g]	54.0723
Bacteria	Chloroflexi	Anaerolinea thermophila UNI-1	atm	54.3740
Bacteria	Proteobacteria	Magnetococcus marinus MC-1	mgm [a]	54.7933
Bacteria	Proteobacteria	Serratia fonticola GS2	sfg [g]	55.0409
Bacteria	Armatimonadetes	Chthonomonas calidirosea T49	ccz	55.1617
Bacteria	Planctomycetes	Rhodopirellula baltica SH 1 (Pirellula sp. strain 1)	rba	55.4576
Bacteria	Proteobacteria	Haematospirillum jordaniae H5569	hjo [a]	55.9861
Bacteria	Proteobacteria	Enterobacter hormaechei subsp. xiangfangensis LMG27195	exf [g]	56.4948
Bacteria	Chrysiogenetes	Desulfurispirillum indicum S5	din	56.7991
Bacteria	Proteobacteria	Serratia plymuthica PRI-2C	sply [g]	57.0459
Archaea	Crenarchaeota	Aeropyrum camini SY1 = JCM 12091	acj	57.3136
Bacteria	Proteobacteria	Desulfohalobium retbaense DSM 5692	drt [d]	57.8989
Bacteria	Proteobacteria	Cronobacter malonaticus CMCC45402 (Cronobacter sakazakii CMCC 45402)	csi [g]	58.1391
Bacteria	Proteobacteria	Klebsiella pneumoniae 30684/NJST258_2	kps [g]	58.2669
Bacteria	Verrucomicrobia	Akkermansia glycaniphila APytT	agl	58.5367
Bacteria	Acidobacteria	Granulicella mallensis MP5ACTX8	gma	58.7453
Bacteria	Proteobacteria	Pseudomonas cichorii JBC1	pci [g]	59.0336
Bacteria	Candidatus Peregrinibacteria	Candidatus Peribacter riflensis	prf	59.2436
Bacteria	Firmicutes	Kyrpidia tusciae DSM 2912	bts	59.5610
Bacteria	Proteobacteria	Dechloromonas aromatica RCB	dar [b]	59.8727
Bacteria	Proteobacteria	Desulfomicrobium orale DSM 12838	doa [d]	60.0706
Bacteria	Proteobacteria	Agrobacterium rhizogenes K599	aro [a]	60.5231
Archaea	Euryarchaeota	Methanoculleus sp. MAB1	mema	60.8900
Bacteria	Proteobacteria	Serratia marcescens WW4	smw [g]	61.0617
Archaea	Euryarchaeota	Methanopyrus kandleri AV19	mka	61.1975
Bacteria	Cyanobacteria	Gloeobacter kilaueensis JS1	glj	61.3675
Bacteria	Proteobacteria	Halothiobacillus sp. LS2	haz [g]	61.5751
Bacteria	Acidobacteria	Chloracidobacterium thermophilum B	ctm	61.8880
Bacteria	Acidobacteria	Candidatus Solibacter usitatus Ellin6076	sus	62.4283
Bacteria	Proteobacteria	Sinorhizobium meliloti Rm41	smi [a]	62.7275
Bacteria	Proteobacteria	Sinorhizobium americanum CFNEI 73	same [a]	63.0290
Bacteria	Proteobacteria	Martelella mediterranea DSM 17316 MACL11	mmed [a]	63.1697
Bacteria	Proteobacteria	Mesorhizobium opportunistum WSM2075	mop [a]	63.5854
Bacteria	Proteobacteria	Chromobacterium rhizoryzae JP2-74	crz [b]	64.0046
Bacteria	Chloroflexi	Thermomicrobium roseum DSM 5159	tro	64.1814
Bacteria	Verrucomicrobia	Opitutaceae bacterium TAV5	obt	64.4881
Bacteria	Proteobacteria	Sulfitobacter sp. AM1-D1	suam [a]	65.0155
Bacteria	Chloroflexi	Candidatus Promineofilum breve Cfx-K	pbf	65.2923
Bacteria	Proteobacteria	Dyella japonica A8	dja [g]	65.5524
Bacteria	Proteobacteria	Novosphingobium resinovorum SA1	nre [a]	65.7624
Bacteria	Proteobacteria	Paracoccus zhejiangensis J6	pzh [a]	66.0626
Bacteria	Proteobacteria	Pseudomonas aeruginosa PA38182	paeu [g]	66.3909
Archaea	Euryarchaeota	Salinigranum rubrum GX10	srub	66.6672
Bacteria	Proteobacteria	Roseateles depolymerans KCTC 42856	rdp [b]	67.0588
Bacteria	Actinobacteria	Rubrobacter radiotolerans RSPS-4	rrd	67.2519
Bacteria	Acidobacteria	Luteitalea pratensis	abac	67.5117
Bacteria	Proteobacteria	Variovorax paradoxus B4	vpd [b]	67.6780
Bacteria	Proteobacteria	Orrella dioscoreae LMG 29303	odi [b]	68.0388
Bacteria	Chloroflexi	Sphaerobacter thermophilus DSM 20745	sti	68.3435
Bacteria	Actinobacteria	Stackebrandtia nassauensis DSM 44728	sna	68.6585
Bacteria	Actinobacteria	Kibdelosporangium phytohabitans KLBMP1111	kphy	68.9564
Bacteria	Cyanobacteria	Cyanobium sp. NIES-981	cyi	69.0785
Bacteria	Firmicutes	Symbiobacterium thermophilum IAM 14863	sth	69.2860
Bacteria	Proteobacteria	Vulgatibacter incomptus DSM 27710	vin [d]	69.4201
Bacteria	Firmicutes	Limnochorda pilosa HC45	lpil	69.9349
Bacteria	Actinobacteria	Alloactinosynnema sp. L-07	alo ●	70.0457
Bacteria	Deinococcus	Deinococcus ficus CC-FR2-10	dfc	70.4350
Bacteria	Actinobacteria	Actinoplanes friuliensis DSM 7358	afs ●	70.8156
Bacteria	Actinobacteria	Streptomyces davaonensis JCM 4913 (Streptomyces davawensis JCM 4913)	sdv ●	71.0467

Tabla S1. Continuación…

Dominio	Filo	Especie	ID_KEGG	%GC genómico
Bacteria	Actinobacteria	Nakamurella multipartita DSM 44233	nml ●	71.2232
Bacteria	Proteobacteria	Sorangium cellulosum So ce56	scl [d]	71.5924
Bacteria	Actinobacteria	Sanguibacter keddieii DSM 10542	ske ●	72.0986
Bacteria	Actinobacteria	Nonomuraea sp. ATCC 55076	noa ●	72.3458
Bacteria	Actinobacteria	Frankia sp. Eul1c	fri ●	72.6910
Bacteria	Actinobacteria	Conexibacter woesei DSM 14684	cwo ●	72.9344
Bacteria	Actinobacteria	Plantactinospora sp. KBS50	plk ●	73.0400
Bacteria	Actinobacteria	Streptomyces cattleya NRRL 8057 = DSM 46488	scy ●	73.2316
Bacteria	Proteobacteria	Anaeromyxobacter sp. Fw109-5	afw [d]	73.5390
Bacteria	Actinobacteria	Cellvibrio gilvus ATCC 13127 (Cellulomonas gilvus ATCC 13127)	cga ●	73.8365
Bacteria	Actinobacteria	Geodermatophilus obscurus DSM 43160	gob ●	74.1259
Bacteria	Actinobacteria	Kineococcus radiotolerans SRS30216	kra ●	74.3394
Bacteria	Actinobacteria	Cellulomonas fimi ATCC 484	cfi ●	74.6389[↑]

● Grupo de Actinobacteria con mayor contenido de GC genómico.

[↓] Candidatus Sulcia muelleri PSPU (smup) es el Bacteroidete endosimbionte con contenido de GC genómico más pequeño en este estudio.

[↑] Cellulomonas fimi ATCC 484 (cfi) es la Actinobacteria con más alto contenido de GC genómico.

[a] Especies bacterianas incluidas dentro de la clase Alphaproteobacteria, ellos presentan un promedio de contenido de GC genómico de 54.41%.

[b] Especies bacterianas incluidas dentro de la clase Betaproteobacteria, ellos presentan una media de contenido de GC genómico de 58.66%.

[d] Especies bacterianas incluidas dentro de la clase Deltaproteobacteria, tienen un promedio de contenido de GC genómico de 57.37%

[g] Especies bacterianas incluidas dentro de la clase Gammaproteobacteria, ellos presentan una media de contenido de GC genómico de 45.53%.

Tabla S2. Regresión lineal de 20 aminoácidos en el proteoma de 192 procariotas con respecto a su contenido de GC genómico. La m (pendiente), b (intercepto), R (correlación), R^2 (coeficiente de determinación) y *p-value* son presentados a partir de la segunda a la sexta columna, respectivamente.

Aminoácido	m	b	R	R^2	*p-value*
A [H]	0.1942	-0.6453	0.9740	0.9487	< 0.001
G [H]	0.0837	3.1541	0.9435	0.8902	< 0.001
P	0.0712	0.8446	0.945	0.893	< 0.001
R [H]	0.1131	-0.0751	0.9545	0.9112	< 0.001
W	0.0164	0.3779	0.8017	0.6427	< 0.001
V	0.0666	3.7122	0.8138	0.6622	< 0.001
H	0.0140	1.2715	0.6139	0.3769	< 0.001
D	0.0147	4.6611	0.3622	0.1312	< 0.001
T	0.0188	4.3333	0.4717	0.2225	< 0.001
Q ***	0.0072	3.0399	0.1255	0.0157	0.0831
C	-0.0024	1.0792	-0.1410	0.0199	0.0519
E	-0.0180	7.1860	-0.2939	0.0864	< 0.001
S	-0.0309	7.4957	-0.6623	0.4387	< 0.001
L	0.0254	8.9142	0.5032	0.2532	< 0.001
M**	-0.0032	2.4477	-0.1257	0.0158	0.0816
F	-0.0467	6.4165	-0.8737	0.7634	< 0.001
Y	-0.0534	5.8465	-0.9054	0.8198	< 0.001
K [L]	-0.1870	14.9165	-0.9590	0.9196	< 0.001
N [L]	-0.1214	10.1984	-0.9506	0.9037	< 0.001
I [L]	-0.1622	14.8264	-0.9558	0.9135	< 0.001

[H] aminoácidos con codones altos en contenido de GC y con valores absolutos de pendiente más altas.
[L] aminoácidos con codones bajos en contenido de GC y con valores absolutos de pendiente más altas.
** = Met en el proteoma se ajusta mejor a una regresión polinomial de orden 2,
$y = -0.0013x^2 + 0.1206x - 0.2143$ con una $R^2 = 0.4648$.
*** = Gln en el proteoma se ajusta mejor a una regresión polinomial de orden 3,
$y = -4E\text{-}05x^3 + 0.0045x^2 - 0.1001x + 3.1997$ con una $R^2 = 0.2178$.

Tabla S3. Regresión lineal de 20 aminoácidos en la estructura secundaria de 192 proteomas con respecto a su contenido de GC genómico. Los datos de m, b, R, R², and *p-value* son presentados para hélices alfa (de la segunda a la sexta columna), hojas beta (a partir de la séptima a la onceava columna) y lazos (a partir de doceava a la decimosexta columna).

ss	Alfa hélice					Beta plegada					Lazo				
Aminoácido	**m**	**b**	**R**	**R²**	***p-value***	**m**	**b**	**R**	**R²**	***p-value***	**m**	**b**	**R**	**R²**	***p-value***
A	0.2549	-1.0588	0.9764	0.9534	< 0.001	0.1150	1.0392	0.9727	0.9462	< 0.001	0.1639	-1.1763	0.9623	0.9261	< 0.001
G	0.0608	1.1557	0.9371	0.8782	< 0.001	0.0275	2.4659	0.8364	0.6996	< 0.001	0.1162	6.6287	0.9021	0.8137	< 0.001
P	0.0364	0.4430	0.9612	0.9238	< 0.001	0.0238	0.4431	0.9397	0.8830	< 0.001	0.1185	2.0880	0.9383	0.8803	< 0.001
R	0.1303	-0.2981	0.9539	0.9099	< 0.001	0.1039	-0.1822	0.9639	0.9292	< 0.001	0.0987	0.1640	0.9419	0.8872	< 0.001
W	0.0189	0.4129	0.7859	0.6177	< 0.001	0.0232	0.3234	0.8154	0.6649	< 0.001	0.0116	0.3145	0.7857	0.6173	< 0.001
V	0.0717	3.0501	0.8280	0.6855	< 0.001	0.1522	7.0687	0.8304	0.6896	< 0.001	0.0349	2.5419	0.7823	0.6121	< 0.001
H	0.0135	1.0495	0.6113	0.3737	< 0.001	0.0169	1.1402	0.7232	0.5231	< 0.001	0.0126	1.6268	0.5102	0.2603	< 0.001
D	0.0094	3.7357	0.2815	0.0793	< 0.001	0.0070	2.6965	0.3628	0.1316	< 0.001	0.0153	7.0147	0.2578	0.0665	< 0.001
T	0.0120	3.7068	0.3718	0.1382	< 0.001	0.0337	4.7445	0.5966	0.3560	< 0.001	0.0181	4.9897	0.4138	0.1713	< 0.001
Q***	0.0087	3.5658	0.1270	0.0161	0.0794	0.0099	2.0761	0.2375	0.0564	< 0.001	0.0059	2.7649	0.1124	0.0126	0.1211
C	-0.0023	0.9390	-0.1396	0.0195	0.0519	-0.0025	1.3918	-0.1142	0.0130	0.1132	-0.0026	1.1149	-0.1381	0.0191	0.0562
E	-0.0211	8.4565	-0.2932	0.0860	< 0.001	-0.0053	4.9624	-0.1437	0.0207	0.0467	-0.0171	6.5451	-0.2770	0.0768	< 0.001
S	-0.0422	7.2902	-0.7831	0.6133	< 0.001	-0.0173	5.5095	-0.4763	0.2269	< 0.001	-0.0289	8.8344	-0.5580	0.3113	< 0.001
L	0.0430	11.2351	0.6348	0.4030	< 0.001	0.0566	8.7165	0.6934	0.4808	< 0.001	0.0071	5.5615	0.2356	0.0555	< 0.001
M**	-0.0028	2.6296	-0.0874	0.0076	0.2268	0.0003	1.9860	0.0104	0.0001	0.8929	-0.0050	2.4248	-0.2509	0.0629	< 0.001
F	-0.0557	7.2703	-0.8850	0.7832	< 0.001	-0.0470	7.9721	-0.7548	0.5697	< 0.001	-0.0324	4.5541	-0.8513	0.7246	< 0.001
Y	-0.0566	6.0155	-0.9184	0.8435	< 0.001	-0.0620	7.6860	-0.8296	0.6883	< 0.001	-0.0438	4.7411	-0.9073	0.8231	< 0.001
K	-0.2016	15.9689	-0.9554	0.9128	< 0.001	-0.1283	10.5723	-0.9571	0.9161	< 0.001	-0.1957	15.6069	-0.9562	0.9143	< 0.001
N	-0.1001	8.2248	-0.9486	0.8998	< 0.001	-0.0549	5.0262	-0.9229	0.8517	< 0.001	-0.1814	15.0854	-0.9527	0.9076	< 0.001
I	-0.1773	16.2073	-0.9591	0.9198	< 0.001	-0.2527	24.3623	-0.9492	0.9009	< 0.001	-0.0958	8.5740	-0.9343	0.8729	< 0.001

ss =estructura secundaria, puede ser hélice alfa, hoja beta o lazo.
** =Met se ajusta mejor a una regresión polinomial de orden 2;
para hélice alfa $y = -0.0016x^2 + 0.1532x - 0.7225$ con una $R^2 = 0.4722$,
para hoja beta $y = -0.0013x^2 + 0.1212x - 0.6137$ con una $R^2 = 0.4684$,
para lazo $y = -0.0009x^2 + 0.0832x + 0.5283$ con una $R^2 = 0.4374$.
*** =Gln se ajusta mejor a una regresión polinomial de orden 3;
los datos para hélice alfa $y = -5E\text{-}05x^3 + 0.0054x^2 - 0.1224x + 3.8258$ con una R^2 de 0.2115,
los datos para hoja beta $y = -3E\text{-}05x^3 + 0.0036x^2 - 0.0899x + 2.5725$ con una $R^2 = 0.2363$,
los datos para lazo $y = -4E\text{-}05x^3 + 0.0035x^2 - 0.0604x + 2.4027$ con una $R^2 = 0.2352$.

Tabla S4. Regresión lineal de los Parámetros Conformacionales de los aminoácidos a partir de 192 proteomas con respecto a su contenido de GC genómico. Los datos de m, b, R, R^2, y *p-value* son presentados para hélices alfa (de la segunda a la sexta columna), hojas beta (de la séptima a la onceava columna) y lazos (de la doceava a la decimosexta columna).

	Alfa hélice					Beta plegada					Lazo				
Aminoácido	**m**	**b**	**R**	**R^2**	***p-value***	**m**	**b**	**R**	**R^2**	***p-value***	**m**	**b**	**R**	**R^2**	***p-value***
Ala	0.0006	1.2552	0.3217	0.1035	<0.001	-0.0044	0.9955	-0.8513	0.7247	<0.001	0.0019	0.6678	0.6105	0.3727	<0.001
Gly	0.00207	0.4627	0.6741	0.4543	<0.001	-0.00237	0.6495	-0.7516	0.5649	<0.001	-0.0040	1.9042	-0.7383	0.5452	<0.001
Pro	0.00015	0.5080	0.0794	0.0063	0.2739	-0.00057	0.4027	-0.3352	0.1124	<0.001	-0.00297	1.9810	-0.6417	0.4117	<0.001
Arg	0.0010	1.0572	0.5577	0.3111	<0.001	0.00057	0.8694	0.2466	0.0608	<0.001	-0.00117	0.9776	-0.5949	0.3539	<0.001
Trp	0.00041	1.1108	0.1376	0.0189	0.0569	0.00257	1.1012	0.4180	0.1747	<0.001	-0.00073	0.7866	-0.2146	0.0460	0.0028
Val	0.0014	0.8658	0.6295	0.3963	<0.001	0.0014	2.0106	0.2513	0.0632	<0.001	-0.00087	0.6541	-0.6550	0.4291	<0.001
His	0.0008	0.8326	0.3116	0.0971	<0.001	0.00138	0.9386	0.3993	0.1594	<0.001	-0.00193	1.2462	-0.5250	0.2757	<0.001
Asp	-0.0004	0.8001	-0.2448	0.0599	<0.001	-0.00028	0.5803	-0.1474	0.0217	0.0414	-0.00095	1.4902	-0.3248	0.1055	<0.001
Thr	-0.0006	0.8473	-0.3217	0.1035	<0.001	0.00211	1.1116	0.6199	0.3843	<0.001	-0.00063	1.1493	-0.3785	0.1433	<0.001
Gln	0.0001	1.1701	0.0681	0.0046	0.3490	0.0011	0.7045	0.3027	0.0916	<0.001	-0.00017	0.9078	-0.0776	0.006	0.2847
Cys	-0.0002	0.8637	-0.0352	0.0012	0.6285	0.00092	1.2866	0.1204	0.0145	0.0961	-0.00022	1.0427	-0.0279	0.0008	0.7006
Glu	0.0000	1.1796	-0.0141	0.0002	0.8456	0.00133	0.6848	0.4747	0.2253	<0.001	-0.00012	0.9086	-0.0537	0.0029	0.4592
Ser	-0.0026	0.9969	-0.803	0.6448	<0.001	0.00111	0.726	0.3739	0.1398	<0.001	0.001629	1.1635	0.624	0.3893	<0.001
Leu	0.0010	1.2660	0.5042	0.2542	<0.001	0.00275	0.9944	0.7337	0.5383	<0.001	-0.00076	0.6192	-0.6006	0.3607	<0.001
Met**	0.00027	1.0709	0.1001	0.01	0.1671	0.00138	0.8039	0.4755	0.2261	<0.001	-0.00082	0.9985	-0.2288	0.0524	0.0014
Phe	-0.0010	1.1481	-0.5594	0.3129	<0.001	0.00433	1.1732	0.7443	0.5541	<0.001	0.000333	0.7034	0.1976	0.0391	0.0060
Tyr	-0.0009	1.0430	-0.5199	0.2703	<0.001	0.00473	1.2261	0.7325	0.7325	<0.001	-0.00029	0.8166	-0.171	0.0292	0.0178
Lys	-0.0007	1.0862	-0.3143	0.0988	<0.001	0.00249	0.6435	0.6783	0.4601	<0.001	-0.00035	1.0591	-0.1582	0.025	0.0285
Asn	-0.0010	0.8259	-0.4352	0.1894	<0.001	0.0028	0.4312	0.7456	0.556	<0.001	-0.00156	1.5242	-0.5493	0.3018	<0.001
Ile	0.0001	1.0892	0.064	0.0041	0.3775	0.00458	1.5438	0.7936	0.6299	<0.001	-0.00048	0.5845	-0.3379	0.1142	<0.001

** =Met se ajusta mejor a una regresión polinomial de orden 2;
para hélice alfa $y = -0.0001x^2 + 0.0109x + 0.8428$ con una $R^2 = 0.3061$,
para hoja beta $y = -6E\text{-}05x^2 + 0.007x + 0.6833$ con una $R^2 = 0.2985$,
para lazo $y = 0.0001x^2 - 0.0152x + 1.3077$ con una $R^2 = 0.3633$.

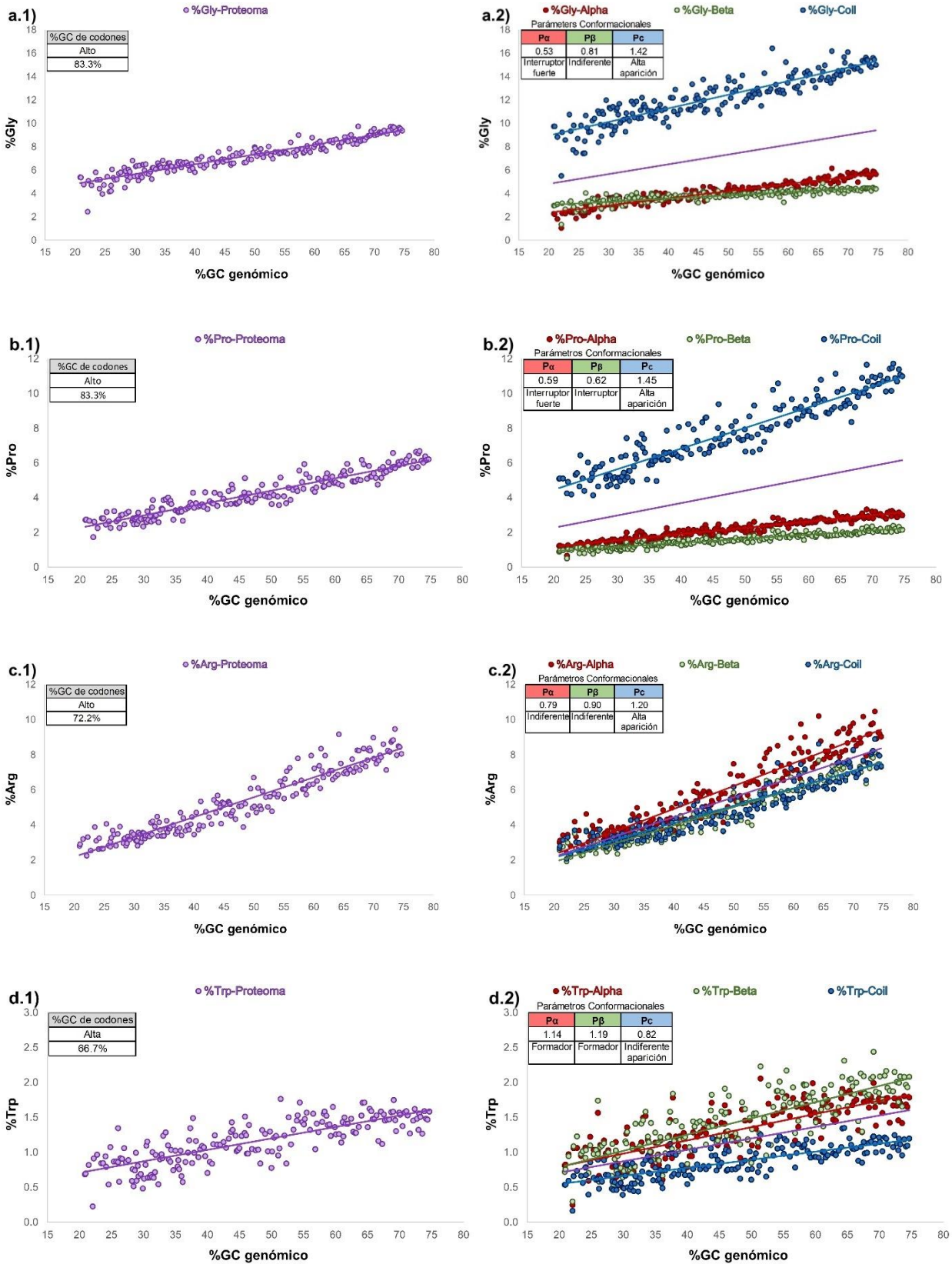
Tabla S5. Regresión lineal de las estructuras secundarias de 192 proteomas con respecto a su contenido de GC genómico. Los datos de m, b, R, R^2 y *p-value* para hélice alfa, hoja beta y lazo son presentadas.

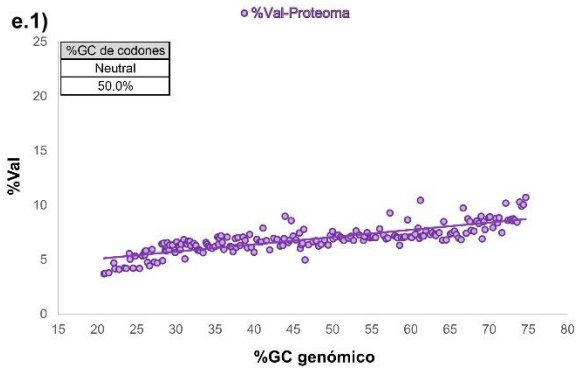
Estructura secundaria	**m**	**b**	**R**	**R^2**	***p-value***
Alfa hélice	-0.0626	47.9061	-0.3602	0.1297	< 0.001
Beta plegada	-0.0222	17.6894	-0.2745	0.0754	< 0.001
Lazo	0.0849	34.4046	0.5900	0.3481	< 0.001

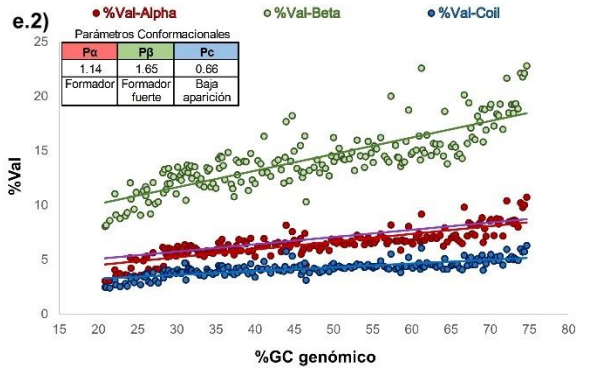
Tabla S6. Regresión lineal de las estructuras secundarias de las proteínas en el COG0002 y COG3228 con respecto al contenido de GC de sus genes. Los datos de m, b, R, R^2 y *p-value* para hélice alfa, hoja beta y lazo en COG0002 (sin sesgo) y COG3228 (con sesgo) son presentadas.

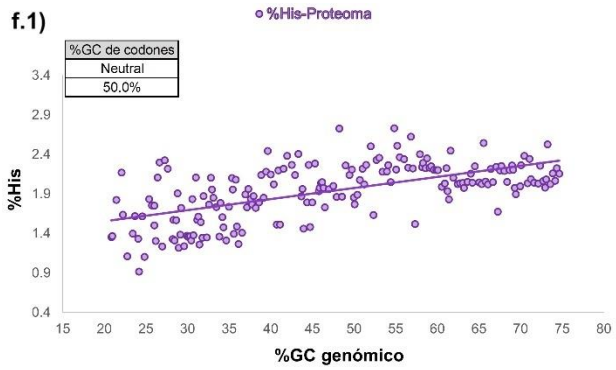
	COG0002					COG3228				
Estructura secundaria	**m**	**b**	**R**	**R^2**	***p-value***	**m**	**b**	**R**	**R^2**	***p-value***
Alfa hélice	-0.0064	33.2253	-0.0668	0.0045	0.0758	-0.1321	61.044	-0.5251	0.2757	<0.001
Beta plegada	0.0045	20.7689	0.0807	0.0065	0.0318	0.0062	9.9318	0.0705	0.0050	0.2780
Lazo	0.0019	45.9894	0.0179	0.0003	0.6348	0.1258	29.011	0.4958	0.2458	<0.001

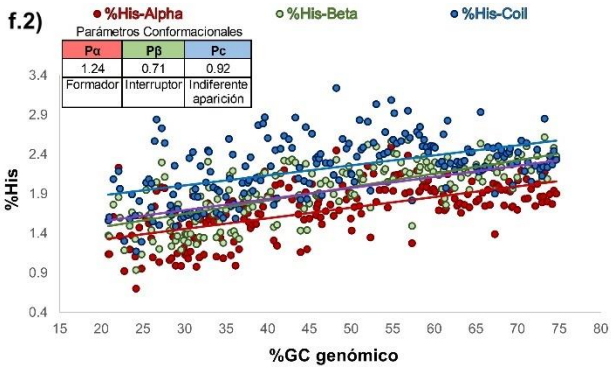
Figuras Complementarias

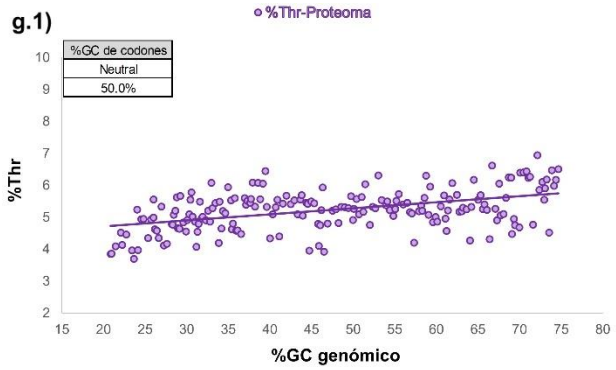


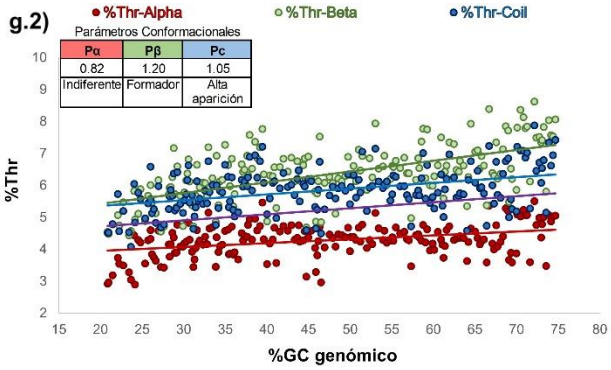

e.1)
%Val-Proteoma
%GC de codones
Neutral
50.0%
%Val
%GC genómico

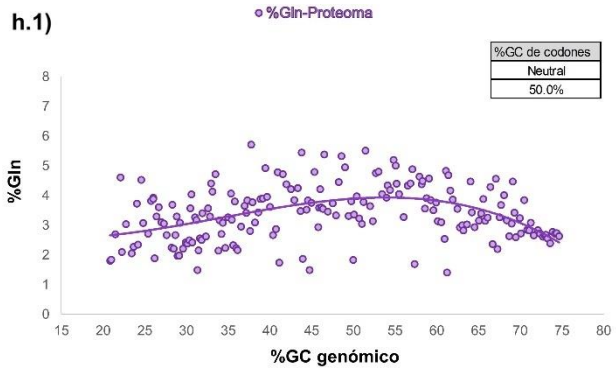

e.2)
%Val-Alpha
%Val-Beta
%Val-Coil
Parámetros Conformacionales
Pα
Pβ
Pc
1.14
1.65
0.66
Formador
Formador fuerte
Baja aparición
%Val
%GC genómico

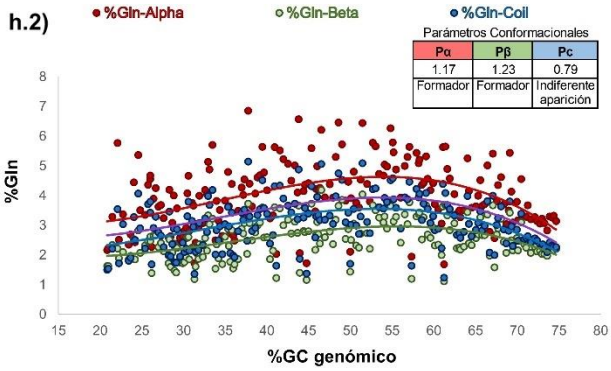

f.1)
%His-Proteoma
%GC de codones
Neutral
50.0%
%His
%GC genómico

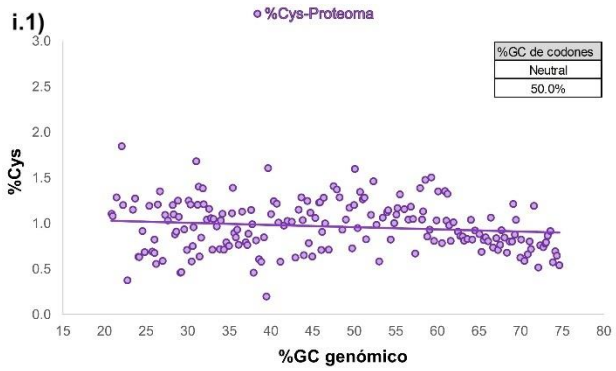

f.2)
%His-Alpha
%His-Beta
%His-Coil
Parámetros Conformacionales
Pα
Pβ
Pc
1.24
0.71
0.92
Formador
Interruptor
Indiferente aparición
%His
%GC genómico

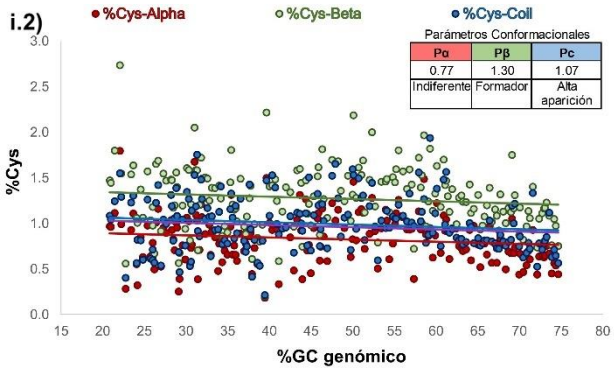

g.1)
%Thr-Proteoma
%GC de codones
Neutral
50.0%
%Thr
%GC genómico

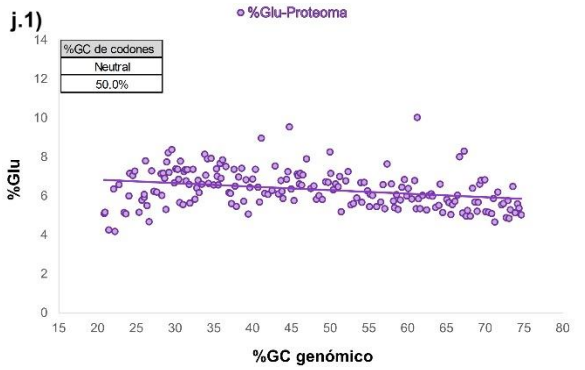

g.2)
%Thr-Alpha
%Thr-Beta
%Thr-Coil
Parámetros Conformacionales
Pα
Pβ
Pc
0.82
1.20
1.05
Indiferente
Formador
Alta aparición
%Thr
%GC genómico

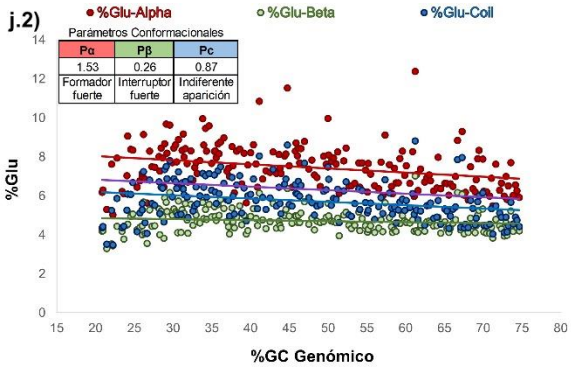

h.1)
%Gln-Proteoma
%GC de codones
Neutral
50.0%
%Gln
%GC genómico

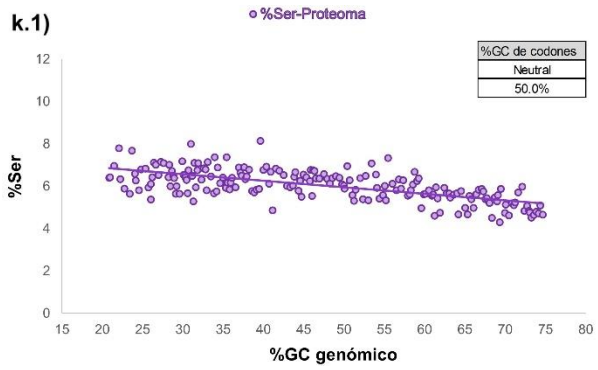

h.2)
%Gln-Alpha
%Gln-Beta
%Gln-Coil
Parámetros Conformacionales
Pα
Pβ
Pc
1.17
1.23
0.79
Formador
Formador
Indiferente aparición
%Gln
%GC genómico

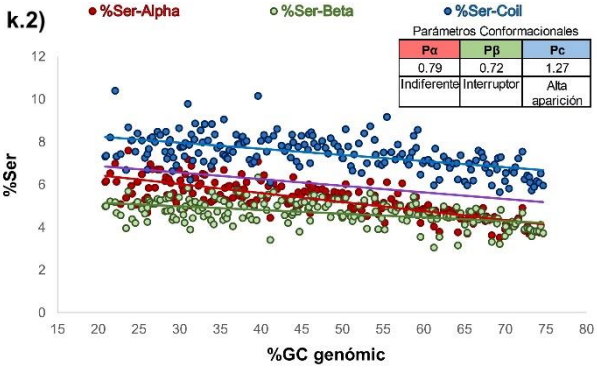

i.1)
%Cys-Proteoma
%GC de codones
Neutral
50.0%
%Cys
%GC genómico

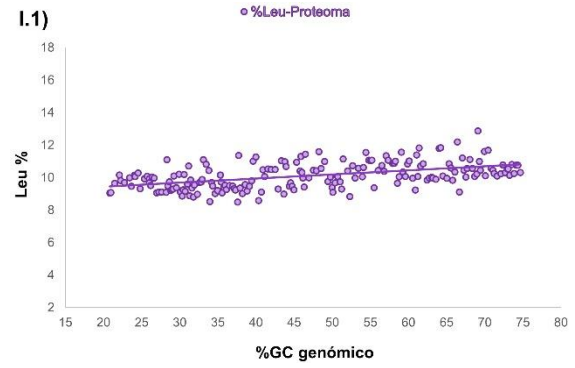

i.2)
%Cys-Alpha
%Cys-Beta
%Cys-Coil
Parámetros Conformacionales
Pα
Pβ
Pc
0.77
1.30
1.07
Indiferente
Formador
Alta aparición
%Cys
%GC genómico

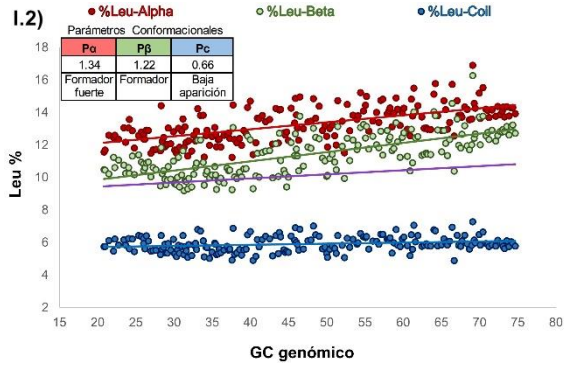

j.1)
%Glu-Proteoma
%GC de codones
Neutral
50.0%
%Glu
%GC genómico

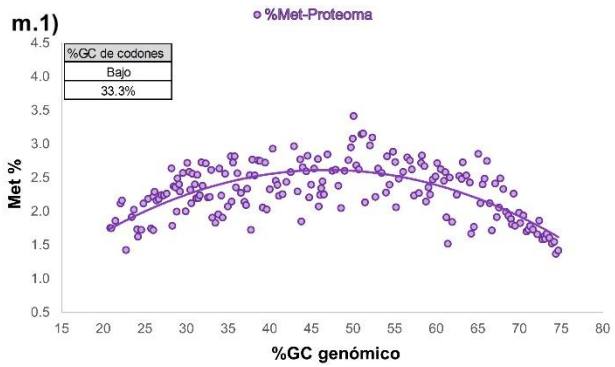

j.2)
%Glu-Alpha
%Glu-Beta
%Glu-Coil
Parámetros Conformacionales
Pα
Pβ
Pc
1.53
0.26
0.87
Formador fuerte
Interruptor fuerte
Indiferente aparición
%Glu
%GC Genómico

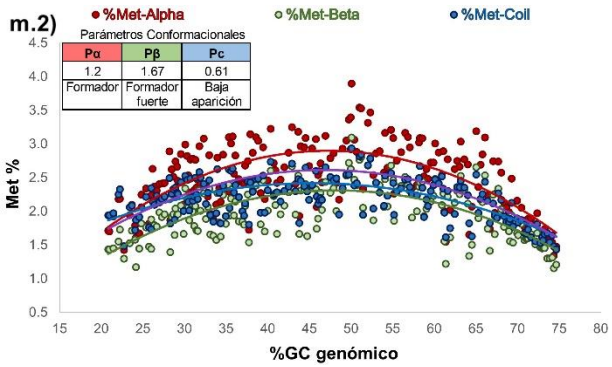

k.1)
%Ser-Proteoma
%GC de codones
Neutral
50.0%
%Ser
%GC genómico

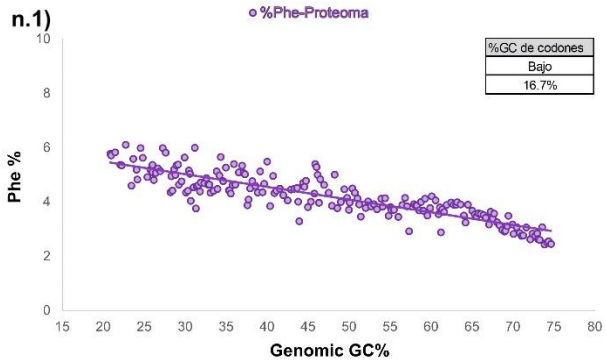

k.2)
%Ser-Alpha
%Ser-Beta
%Ser-Coil
Parámetros Conformacionales
Pα
Pβ
Pc
0.79
0.72
1.27
Indiferente
Interruptor
Alta aparición
%Ser
%GC genómic

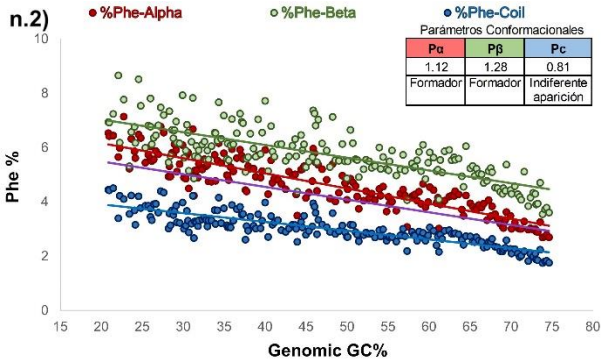

l.1)
%Leu-Proteoma
Leu %
%GC genómico

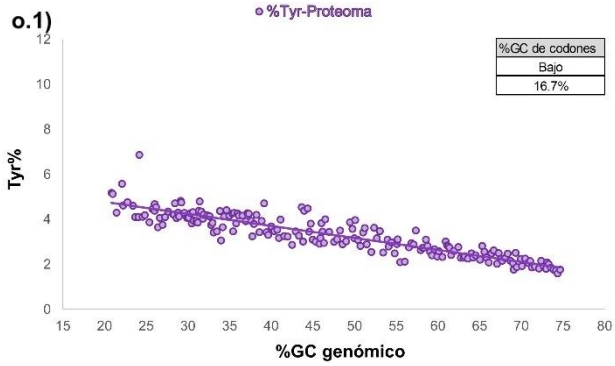

l.2)
%Leu-Alpha
%Leu-Beta
%Leu-Coil
Parámetros Conformacionales
Pα
Pβ
Pc
1.34
1.22
0.66
Formador fuerte
Formador
Baja aparición
Leu %
GC genómico

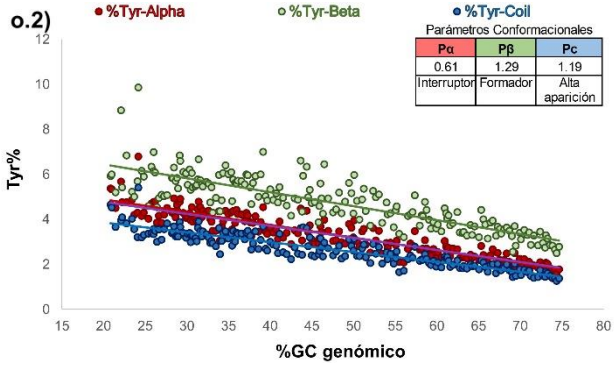

m.1)
%Met-Proteoma
%GC de codones
Bajo
33.3%
Met %
%GC genómico

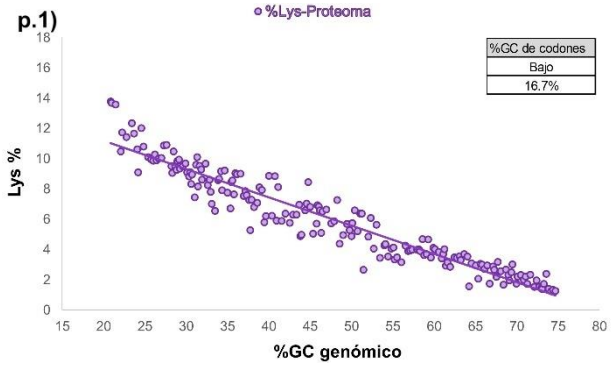

m.2)
%Met-Alpha
%Met-Beta
%Met-Coil
Parámetros Conformacionales
Pα
Pβ
Pc
1.2
1.67
0.61
Formador
Formador fuerte
Baja aparición
Met %
%GC genómico

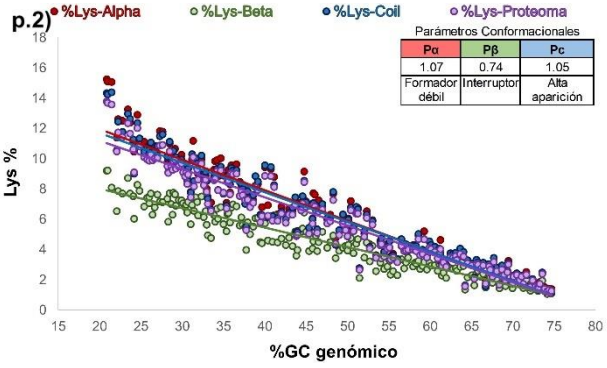

n.1)
%Phe-Proteoma
%GC de codones
Bajo
16.7%
Phe %
Genomic GC%

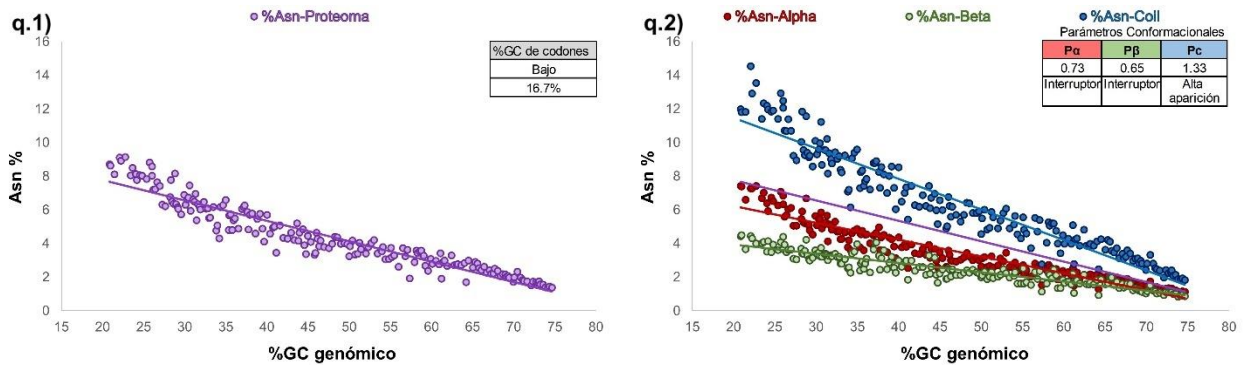

n.2)
%Phe-Alpha
%Phe-Beta
%Phe-Coil
Parámetros Conformacionales
Pα
Pβ
Pc
1.12
1.28
0.81
Formador
Formador
Indiferente aparición
Phe %
Genomic GC%


o.1)
%Tyr-Proteoma
%GC de codones
Bajo
16.7%
Tyr%
%GC genómico

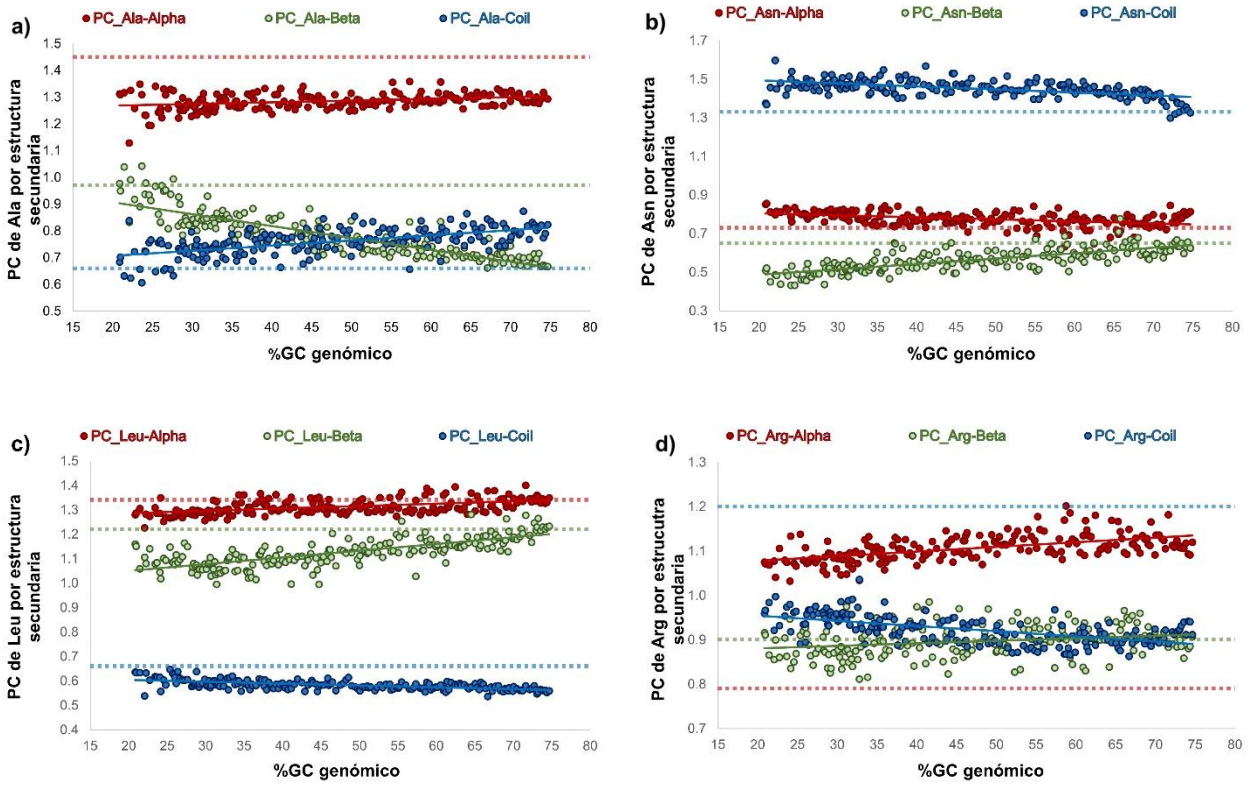

o.2)
%Tyr-Alpha
%Tyr-Beta
%Tyr-Coil
Parámetros Conformacionales
Pα
Pβ
Pc
0.61
1.29
1.19
Interruptor
Formador
Alta aparición
Tyr%
%GC genómico

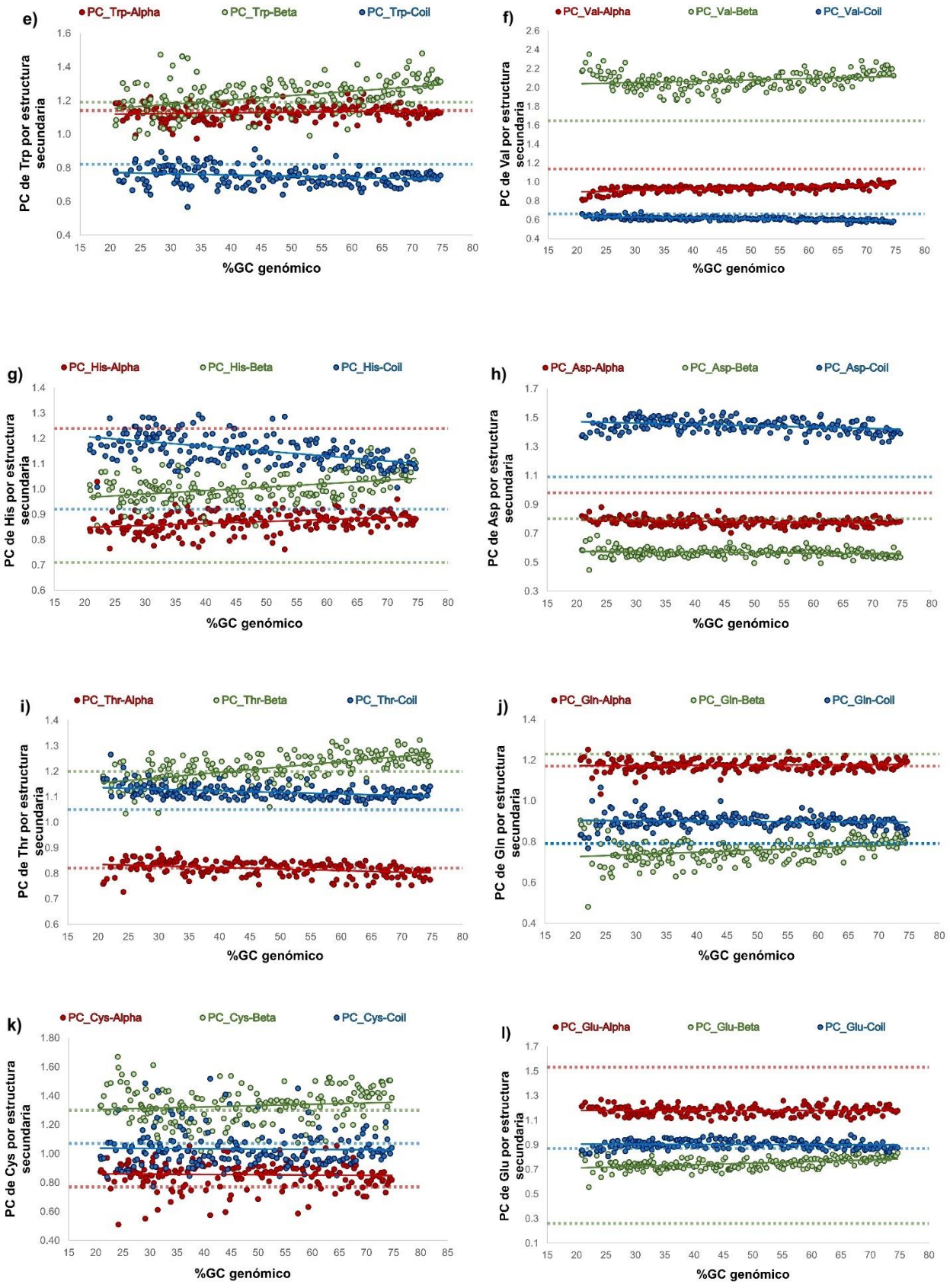

p.1)
%Lys-Proteoma
%GC de codones
Bajo
16.7%
Lys %
%GC genómico

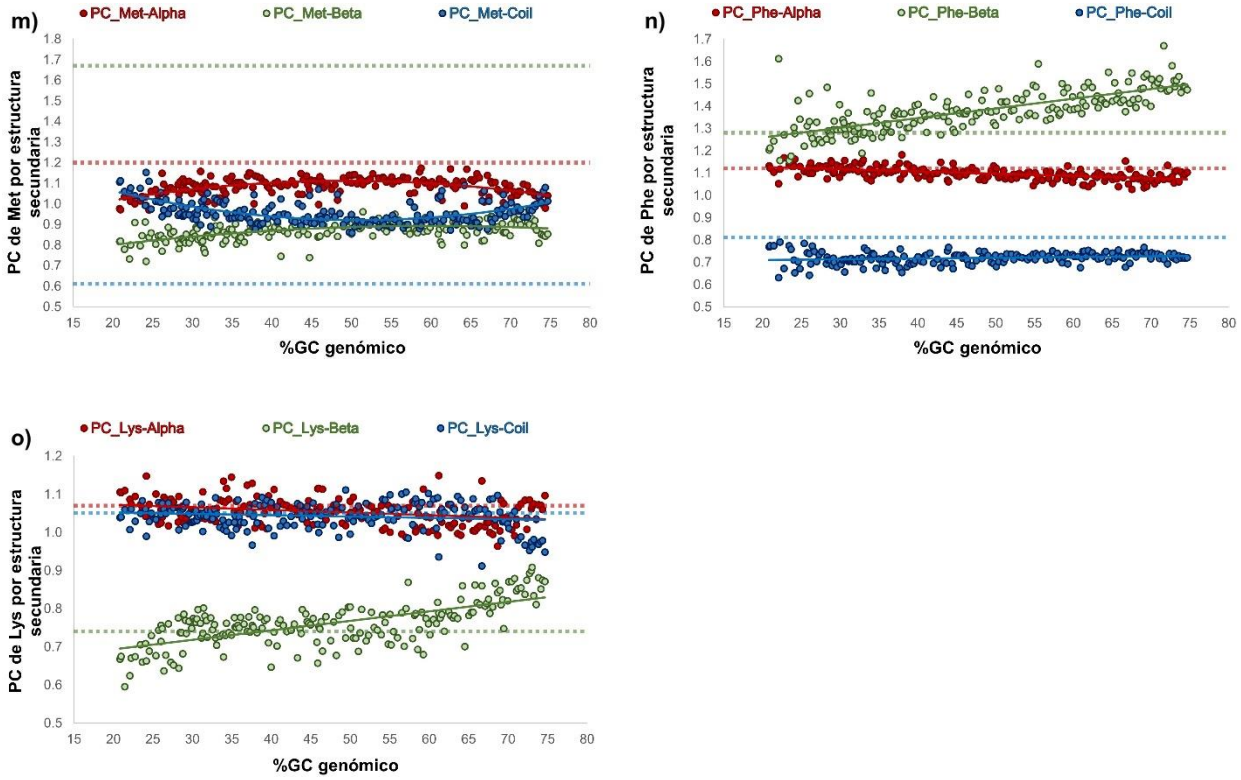

p.2)
%Lys-Alpha
%Lys-Beta
%Lys-Coil
%Lys-Proteoma
Parámetros Conformacionales
Pα
Pβ
Pc
1.07
0.74
1.05
Formador débil
Interruptor
Alta aparición
Lys %
%GC genómico



S1 Figure. Sesgo del contenido de GC genómico sobre las frecuencias de aminoácidos del proteoma y sobre las estructuras secundarias de las proteínas. La frecuencia de los aminoácidos del proteoma, así como el contenido de GC de sus codones (a.1-q.1) y la frecuencia de los aminoácidos en las estructuras secundarias de las proteínas, así como los parámetros conformacionales descritos por Chou y Fasman[2,3] (a.2-q.2) son presentados.




e)
PC_Trp-Alpha
PC_Trp-Beta
PC_Trp-Coil
PC de Trp por estructura secundaria
%GC genómico
f)
PC_Val-Alpha
PC_Val-Beta
PC_Val-Coil
PC de Val por estructura secundaria
%GC genómico
g)
PC_His-Alpha
PC_His-Beta
PC_His-Coil
PC de His por estructura secundaria
%GC genómico
h)
PC_Asp-Alpha
PC_Asp-Beta
PC_Asp-Coil
PC de Asp por estructura secundaria
%GC genómico
i)
PC_Thr-Alpha
PC_Thr-Beta
PC_Thr-Coil
PC de Thr por estructura secundaria
%GC genómico
j)
PC_Gln-Alpha
PC_Gln-Beta
PC_Gln-Coil
PC de Gln por estructura secundaria
%GC genómico
k)
PC_Cys-Alpha
PC_Cys-Beta
PC_Cys-Coil
PC de Cys por estructura secundaria
%GC genómico
l)
PC_Glu-Alpha
PC_Glu-Beta
PC_Glu-Coil
PC de Glu por estructura secundaria
%GC genómico



S2 Figure. Sesgo del contenido de GC genómico sobre los parámetros conformacionales (PC) de los aminoácidos por estructura secundaria de los proteomas. Los valores de PC de los aminoácidos en hélice alfa (círculos rojos), hoja beta (círculos verdes) y lazos (círculos azules) son evaluados en función al contenido de GC genómico. Las líneas de regresión de las estructuras secundarias son representadas por líneas sólidas, mientras que los valores de PC reportados por Chou y Fasman[2,3] son representados por líneas punteadas (en rojo para hélice alfa, en verde para hoja beta y azul para lazo).

N-terminal

```
20_smg-SMGWSS_110           ----------CEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
26_bhy-BHWA1_00539          ----------CEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
27_bapu-BUMPUSDA_CDS00541   ----------CCEEEEECCCCHHHHHHHHHHHHCCCCEEEEEEEECCCCCCCEHH---H
28_cdf-CD630_20340          --------CCCEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
29_icp-ICMP_025             ----------CEEEEEECCCCHHHHHHHHHHHHCCCCEEEEEEECCCCCCCCEHH---H
30_dtn-DTL3_1460            ----------CEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
31_amar-AMRN_0825           ----------CEEEEECCCCHHHHHHHHHHHHCCCCEEEEEE----CCCCCCEHH---H
32_erg-ERGA_CDS_08180       -------CCCCEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHHHHCH
33_csr-Cspa_c07990          ----------CEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHHHHCH
34_elv-FNIIJ_111            ----------CCEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEC---CCCCCCEHH---H
35_aman-B6F84_08265         ----------CCEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
36_apib-G8C43_08245         --------CCCEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
37_cst-CLOST_0139           ----------CCEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHHHHCH
38_bths-CNY62_02780         ----------CEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
39_iag-Igag_1754            --------CCCEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
40_acd-AOLE_08265           ----------CEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEC---CCCCCEEHH---H
41_aar-Acear_1552           ----------CEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
42_cdiv-CPM_0548            ----------CEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
43_enn-FRE64_00255          -----CCCCCCEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
44_aalg-AREALGSMS7_00808    ----------CCEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
45_bcl-ABC2558              ----------CEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
46_acy-Anacy_2677           -----CCCCCEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
47_ava-Ava_3516             -----CCCCCCEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
48_bif-N288_07475           ----------CEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCE-----E
49_elim-B2M23_17620         ----------CEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEC---CCCCCCEHH---H
50_bbe-BBR47_53030          ----------CEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
51_amr-AM1_0609             -----CCCCCCEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCEEHH---H
52_alr-DS731_18710          ----------CEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEECCCCCCCCCHHHHCH
53_atm-ANT_12400            --------CCEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
54_caby-Cabys_3149          --------CCCEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
55_ahn-NCTC12129_04811      ----------CCEEEEECCCCHHHHHHHHHHHHCCCCEEEEEEECCCCCCCCEEHH---H
56_enc-ECL_05030            ----------CEEEEEECCCCHHHHHHHHHHHHCCCCEEEEEEECCCCCCCCEEHH---H
57_aqg-HRU87_04435          ---------CCEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
58_aprs-BI364_03645         --CCCCCCCCEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
59_cag-Cagg_3596            ---------CCEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCEEHH---H
60_cap-CLDAP_16910          ----------CCEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEC---CCCCCEEHH---H
61_adg-Adeg_1833            ----------CEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
62_ddt-AAY81_01845          ---------CCEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEC---CCCCCEEHH---H
63_ahy-AHML_03040           ----------CEEEEEECCCCHHHHHHHHHHHHCCCCEEEEEEECCCCCCCCEHH---H
64_max-MMALV_06140          ---------CCEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCCHH---H
65_boh-AKI39_02980          ---CCCCCCCCEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
66_axy-AXYL_00663           ---CCCCCCCCEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCEEHH---H
67_nbg-DV706_07695          -----CCCCCCEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
68_azl-AZL_a00330           CCCCCCCCCCCCEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHHHHCH
69_abac-LuPra_00731         ---------------CCCCCCHHHHHHHHHHHHCCCEEEEEEEEE--CCCCCC------E
70_ach-Achl_1497            ---------CCEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
71_aey-CDG81_09890          ---------CCEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
72_amyy-YIM_31495           -------CCCCEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
73_brx-BH708_07090          -------C-CCEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
74_ank-AnaeK_0180           -----CCCCCCEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
75_agg-C1N71_09310          ---------CCEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
76_atl-Athai_30770          ---------CCEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEE---CCCCCCEHH---H
77_mcab-HXZ27_20325         ---------CCEEEEEECCCCHHHHHHHHHHHHCCCEEEEEEEC---CCCCCCCHH---H
```

Fig S3. Continuación…

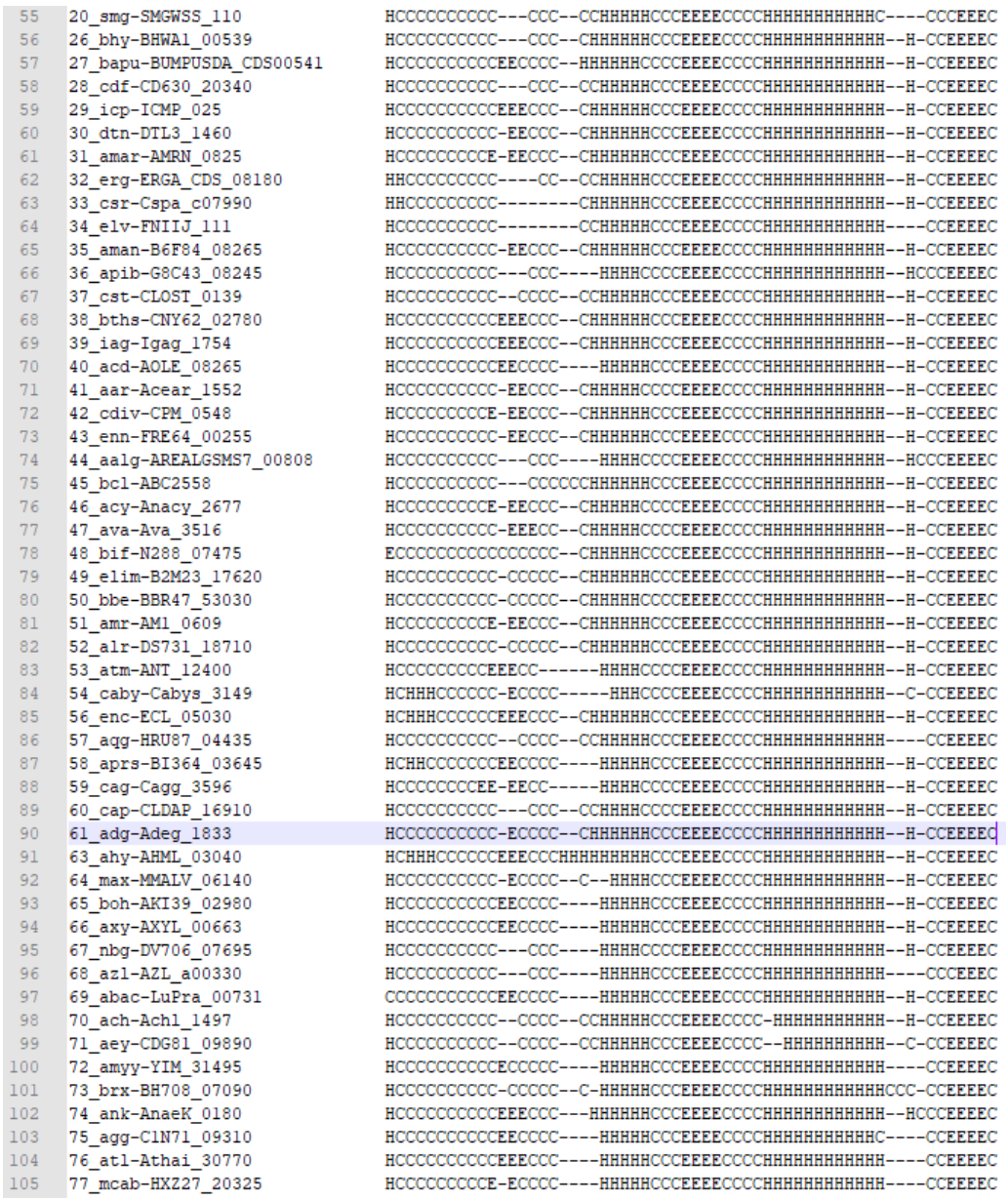


Fig S3. Continuación…

```
20_smg-SMGWSS_110            CCCCCCCCCHHHHHHHHCCCCCC-------------CCC-----------------HHHHH
26_bhy-BHWA1_00539           CCCCCCCCCHHHHHHHHCCCCCCHHHHHHCCCECCCCCC---CCC--------CHHHHHHH
27_bapu-BUMPUSDA_CDS00541    CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------HHHHHHHH
28_cdf-CD630_20340           CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------HHHHHHHH
29_icp-ICMP_025              CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------HHHHHHHH
30_dtn-DTL3_1460             CCCCCCCCCHHHHHHHHCCCCCCCHHHHC-------CCC---CCCC-------CCCHHHHH
31_amar-AMRN_0825            CCCCCCC--HHHHHHHHCCCCCHHHHHH-------CCC---CCC--------CCCHHHHH
32_erg-ERGA_CDS_08180        CCCCCCCCCHHHHHHHHCCCCCHHHHCC--------CCC---CCC--------CCCHHHHH
33_csr-Cspa_c07990           CCCCCCCCCHHHHHHHHCCCCCCCHHHCC-------CCC---CCC--------HHHHHHHH
34_elv-FNIIJ_111             CCCCCCCCCHHHHHHCCCCCCC-------------CCC-----------------HHHHH
35_aman-B6F84_08265          CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------CCCHHHHH
36_apib-G8C43_08245          CCCCCCCCCHHHHHHCCCCCCC-------------CCC-----------------HHHHH
37_cst-CLOST_0139            CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------HHCHHHHH
38_bths-CNY62_02780          CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------HHCHHHHH
39_iag-Igag_1754             CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------CCCHHHHH
40_acd-AOLE_08265            CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------CHCHHHHH
41_aar-Acear_1552            CCCCCCCCCHHHHHHHHCCCCCCHHHHHC-------CCC---CCC--------CCCHHHHH
42_cdiv-CPM_0548             CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------CCCHHHHH
43_enn-FRE64_00255           CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------HHCHHHHH
44_aalg-AREALGSMS7_00808     CCCCCCCCCHHHHHH--------------------CCCCEECCC---------CCHHHHH
45_bcl-ABC2558               CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------HHHHHHHH
46_acy-Anacy_2677            CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------HHCHHHHH
47_ava-Ava_3516              CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------HHCHHHHH
48_bif-N288_07475            CCCCCCCCCHHHHHHH-CCCCCCHHHHHH-------CCC---CCC--------HHCHHHHH
49_elim-B2M23_17620          CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------HHCHHHHH
50_bbe-BBR47_53030           CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------CCCHHHHH
51_amr-AM1_0609              CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------HHHHHHHH
52_alr-DS731_18710           CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------HHHHHHHH
53_atm-ANT_12400             CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------HHHHHHHH
54_caby-Cabys_3149           CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------CCCHHHHH
56_enc-ECL_05030             CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------HHHHHHHH
57_aqg-HRU87_04435           CCCCCCCCCHHHHHHHHCCCCCC-------------CCC---CCCCCCCCHHHHHHHHHH
58_aprs-BI364_03645          CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------HHHHHHHH
59_cag-Cagg_3596             CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------CCCHHHHH
60_cap-CLDAP_16910           CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------HHHHHHHH
61_adg-Adeg_1833             CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------HHCHHHHH
63_ahy-AHML_03040            CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------CCCHHHHH
64_max-MMALV_06140           CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------CCCHHHHH
65_boh-AKI39_02980           CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------HHCHHHHH
66_axy-AXYL_00663            CCCCCCCCCHHHHHHHHCCCCCCCHHHHH-------CCC---CCC--------HHHHHHHH
67_nbg-DV706_07695           CCCCCCCCCHHHHHHHHCCCCCHHHHHH--------CCC---CCC--------HHHHHHHH
68_azl-AZL_a00330            CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------HHHHHHHH
69_abac-LuPra_00731          CCCCCCCCCHHHHHHHHCCCC--HHHHHH-------CCC---CCC--------CCCHHHHH
70_ach-Achl_1497             CCCCCCCCCHHHHHHHHCCCCCC-------------CCCEEECCC---------CCHHHHH
71_aey-CDG81_09890           CCCCCCCCCHHHHHHHHCCCCCC-------------CCCCCCCCC--------CHHHHHHH
72_amyy-YIM_31495            CCCCCCCCCHHHHHHHHCCCCCC-------------CCCEEECCC---------CHHHHHHH
73_brx-BH708_07090           CCCCCCCCCHHHHHHHHCCCCCC-------------CCC---CECCCCCCCHHHHHHHHHH
74_ank-AnaeK_0180            CCCCCCCCCHHHHHHHHCCCCCCHHHHHH-------CCC---CCC--------CCCHHHHH
75_agg-C1N71_09310           CCCCCCCCCHHHHHHHHCCCCCC-------------CCCEECCCCCCCC----CHHHHHHH
76_atl-Athai_30770           CCCCCCCCCHHHHHHHHCCCCCC-------------CCCCCCCCC--------CHHHHHHH
77_mcab-HXZ27_20325          CCCCCCCCCHHHHHHHHCCCCCC-------------CCCCCCCCC--------CHHHHHHH
```

Fig S3. Continuación…

```
159  20_smg-SMGWSS_110            CCCEEECCCCHHHHHHHHHHHHHHH---CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
160  26_bhy-BHWA1_00539           HCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHHC
161  27_bapu-BUMPUSDA_CDS00541    CCCEEECCCCHHHHHHHHHHHHHHH-CCCCCCCCEEEEEECCCCCCCCCCCC------C
162  28_cdf-CD630_20340           CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
163  29_icp-ICMP_025              CCCEEECCCCHHHHHHHHHHHHHHH-CCCCCCCCEEEEEECCCCCCCCCCCCCCC---C
164  30_dtn-DTL3_1460             CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHHC
165  31_amar-AMRN_0825            CCCEEECCCCHHHHHHHHHHHHHHH---CCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
166  32_erg-ERGA_CDS_08180        CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
167  33_csr-Cspa_c07990           CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCCCCC
168  34_elv-FNIIJ_111             CCCEEECCCCHHHHHHHHHHHHHHH---CCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
169  35_aman-B6F84_08265          CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
170  36_apib-G8C43_08245          CCCEEECCCCHHHHHHHHHHHHHHH---CCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
171  37_cst-CLOST_0139            CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
172  38_bths-CNY62_02780          CCCEEECCCCHHHHHHHHHHHHHHH----CCCCCCEEEEEECCCCCCCCCCCCCCCCCC
173  39_iag-Igag_1754             CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCHHCHHHC
174  40_acd-AOLE_08265            CCCEEECCCCHHHHHHHHHHHHHHHCCCCCCCCEEEEEECCCCCCCCCCCCCCC---C
175  41_aar-Acear_1552            CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
176  42_cdiv-CPM_0548             CCCEEECCCCHHHHHHHHHHHHHHH---CCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
177  43_enn-FRE64_00255           CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
178  44_aalg-AREALGSMS7_00808     CCCEEECCCCHHHHHHHHHHHHHHH---CCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
179  45_bcl-ABC2558               CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCCCCC
180  46_acy-Anacy_2677            CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
181  47_ava-Ava_3516              CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
182  48_bif-N288_07475            CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
183  49_elim-B2M23_17620          CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEEECCCCCCCCCCCCCCHHHC
184  50_bbe-BBR47_53030           CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
185  51_amr-AM1_0609              CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
186  52_alr-DS731_18710           CCCEEECCCCHHHHHHHHHHHHHHH-CCCCCCCCEEEEEECCCCCCCCCCCCCCC---C
187  53_atm-ANT_12400             CCCEEECCCCHHHHHHHHHHHHHHH-CCCCCCCCEEEEEEECCCCCCCCCCCCCCHHHC
188  54_caby-Cabys_3149           CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHHC
189  56_enc-ECL_05030             CCCEEECCCCHHHHHHHHHHHHHHH-CCCCCCCCEEEEEECCCCCCCCCCCCCCCC--C
190  57_aqg-HRU87_04435           CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
191  58_aprs-BI364_03645          CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
192  59_cag-Cagg_3596             CCCEEECCCCHHHHHHHHHHHHHHH-CCCCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
193  60_cap-CLDAP_16910           CCCEEECCCCHHHHHHHHHHHHHHH---CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
194  61_adg-Adeg_1833             CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
195  63_ahy-AHML_03040            CCCEEECCCCHHHHHHHHHHHHHHH-CCCCCCCCEEEEEECCCCCCCCCCCCCCC---C
196  64_max-MMALV_06140           CCCEEECCCCHHHHHHHHHHHHHHH---CCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
197  65_boh-AKI39_02980           CCCEEECCCCHHHHHHHHHHHHHHHCCCCCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
198  66_axy-AXYL_00663            CCCEEECCCCHHHHHHHHHHHHHHHCCCCCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
199  67_nbg-DV706_07695           CCCEEECCCCHHHHHHHHHHHHHHH-CCCCCCCCEEEEEECCCCCCCCCCCCCCCHHHH
200  68_azl-AZL_a00330            CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEEECCCCCCCCCCCCCCCHHHC
201  69_abac-LuPra_00731          CCCEEECCCCHHHHHHHHHHHHHHH---CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
202  70_ach-Achl_1497             CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHHH
203  71_aey-CDG81_09890           CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHHC
204  72_amyy-YIM_31495            CCCEEECCCCHHHHHHHHHHHHHHH---CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHC
205  73_brx-BH708_07090           CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHHC
206  74_ank-AnaeK_0180            CCCEEECCCCHHHHHHHHHHHHHHH---CCCCCCEEEEEECCCCCCCCCCCCCCCHHHHC
207  75_agg-C1N71_09310           CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEECCCCCCCCCCCCCCCHHHH
208  76_atl-Athai_30770           CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEEECCCCCCCCCCCCCCHHHH
209  77_mcab-HXZ27_20325          CCCEEECCCCHHHHHHHHHHHHHHH--CCCCCCCEEEEEEECCCCCCCCCCCCCCHHHC
```

Fig S3. Continuación…

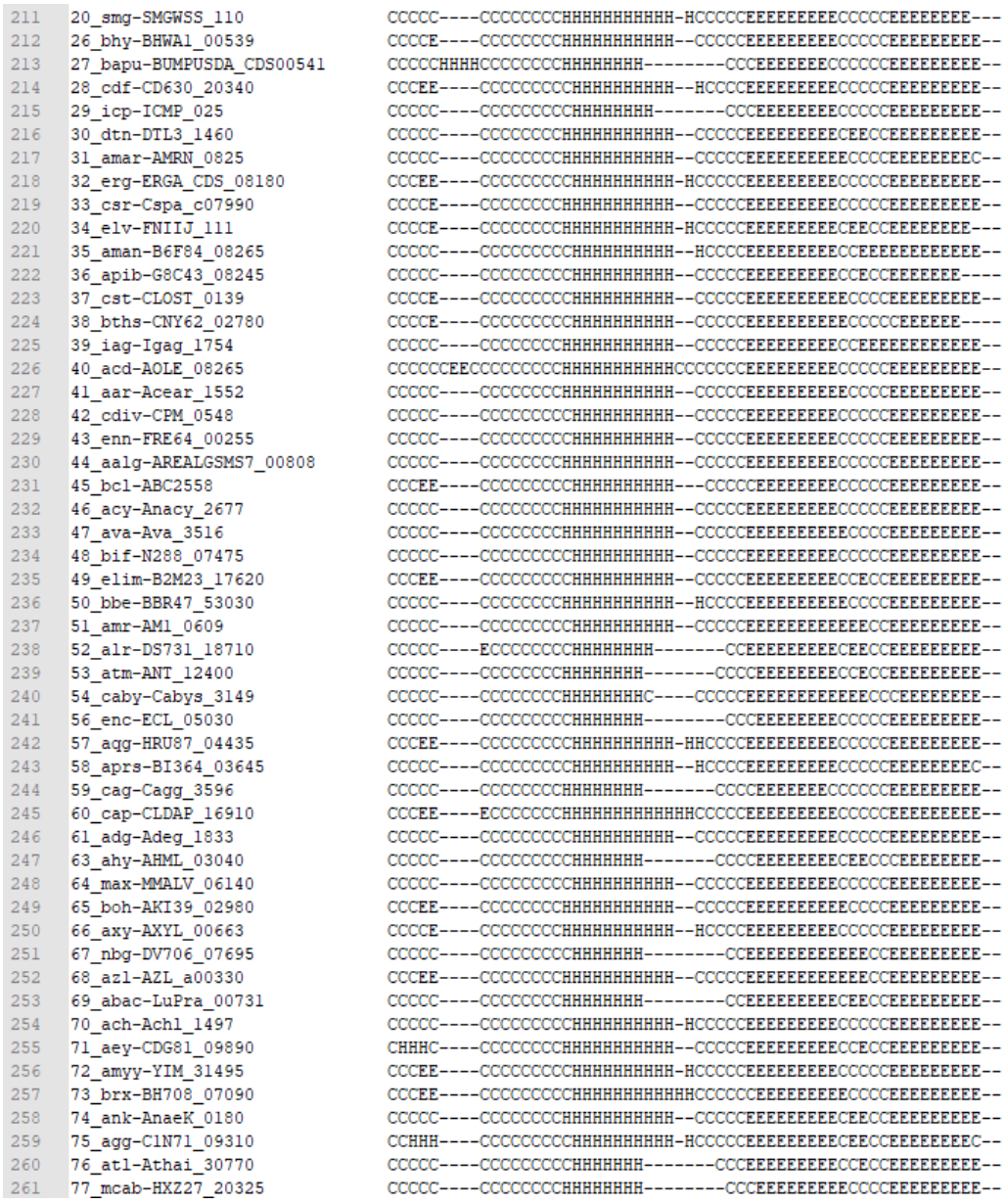


Fig S3. Continuación…

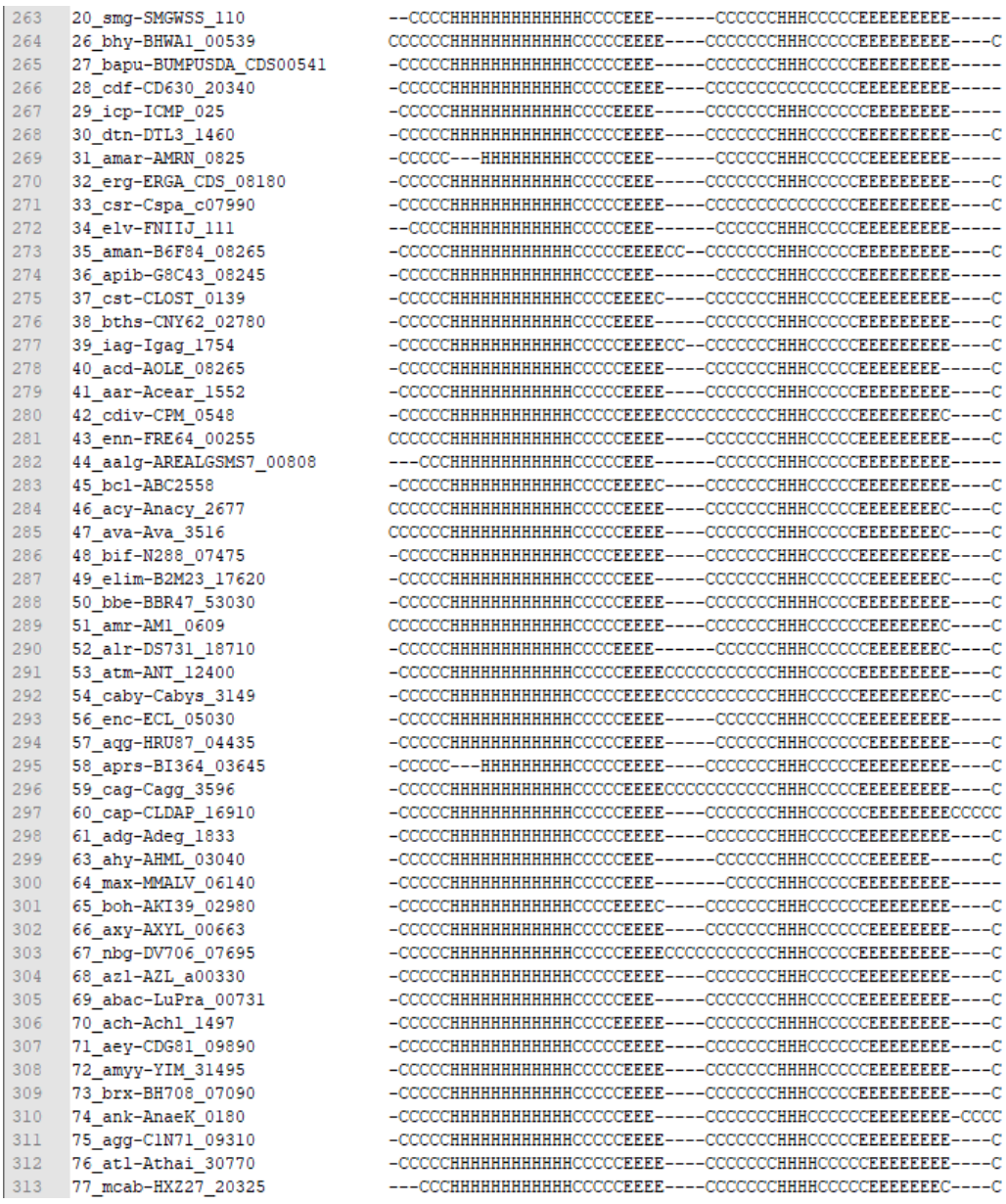
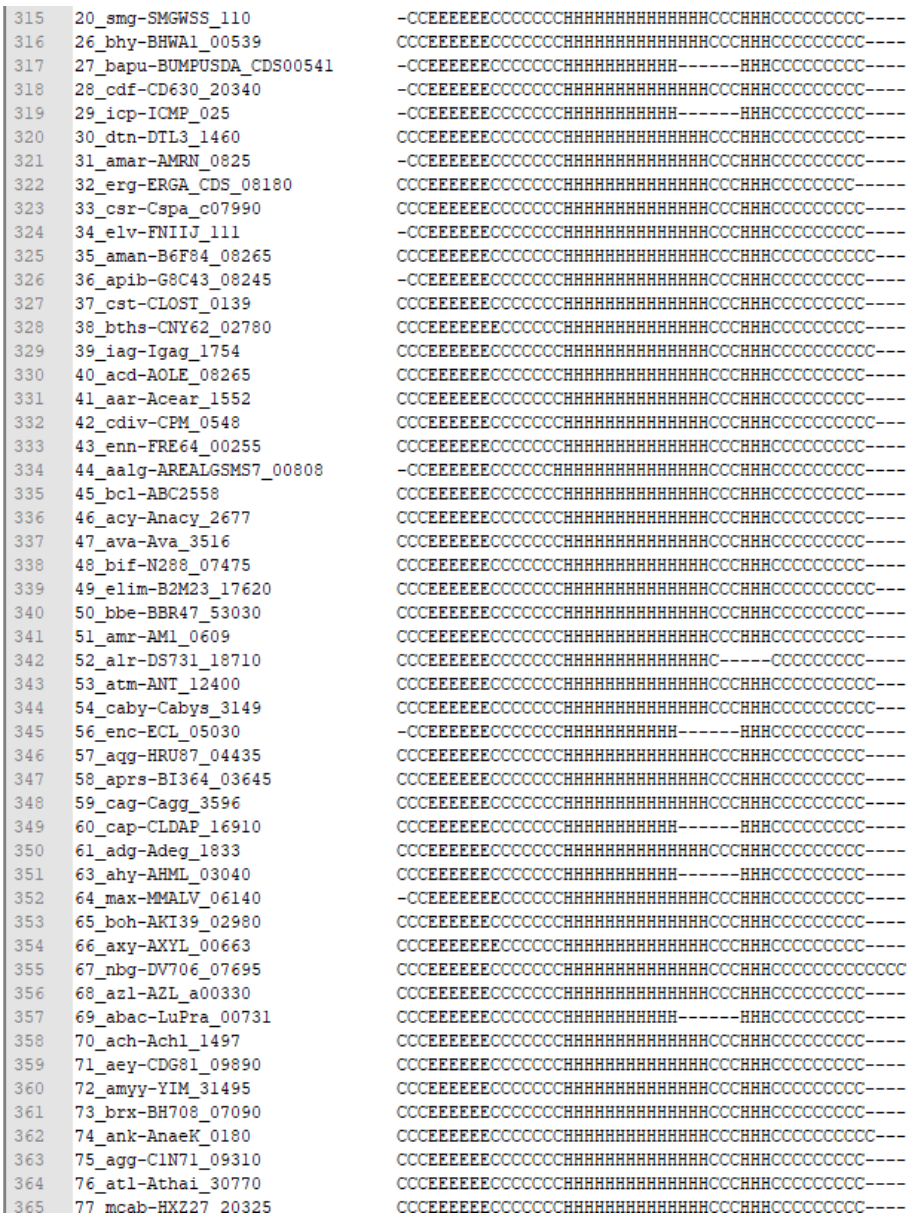


Fig S3. Continuación…



C-terminal

Fig S3. Alineamiento múltiple de los elementos de estructura secundaria de las proteínas del COG0002 con respecto al contenido de GC de sus genes. El formato de alineamiento se divide en dos partes: la primera parte de lado izquierdo refleja el valor (orden ascendente) del contenido de GC de los genes que codifican a las proteínas del COG, seguido del nombre del gen; la segunda parte muestra el alineamiento de las secuencias de estructuras secundarias: H para hélice alfa, E para hoja beta, y C para lazo.

N-terminal

```
27_tje-TJEJU_2162          ------------------HHHHHHHHHHHHHHHHHHHHHHHHHC--------------
28_prn-BW723_13825         ------------------CHHHHHHHHHHHHHHHHHHHHHHHHHH-------------
30_aqb-D1818_09550         CCCCCCEEEEECCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCC
31_laci-CW733_06300        ------------------CHHHHHHHHHHHHHHHHHHHHHHC------------------
36_cep-Cri9333_2611        ------------------CCHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH-----------
37_ava-Ava_0677            ------------------CCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHH-----------
40_cthe-Chro_1359          ------------------CCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHH-----------
41_ptn-PTRA_a1372          ------------------CHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCC-----------
42_zga-ZOBELLIA_2563       ------------CCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHHHHH------------
45_rhh-E0Z06_01765         ------------------CHHHHHHHHHHHHHHHHHHHHHHHHHHHHH------------
46_pko-PKOR_15130          ------------------CHHHHHHHHHHHHHHHHHHHHHHHHHHHHC------------
49_nii-Nit79A3_3401        ------------------CCHHHHHHHHHHH-----------------------------
50_ifl-C1H71_04530         ------------------CHHHHHHHHHHHHHHHHHHHHHHHH----------------
53_mpsy-CEK71_12595        ------------------CCHHHHHHHHHHHH----------------------------
57_dal-Dalk_5292           ------------------CCCHHHHHHHHHHH----------------------------
62_pol-Bpro_2870           ------------------CCCHHHHHHHHHHHC----------------------------
63_eba-ebA6530             ------------------CCCHHHHHHHHHHC----------------------------
64_adi-B5T_02407           ------------------CHHHHHHHHHHHHH----------------------------
65_haz-A9404_06910         --------------CCCCCHHHHHHHHHHHH----------------------------
66_bfz-BAU07_18345         ----------CCCCCCCCCHHHHHHHHH--------------------------------
69_csa-Csal_2711           ------------------CCHHHHHHCC---------------------------------
71_pmex-H4W19_15985        ------------------CCCHHHHC-----------------------------------
```

Continuación…

```
27_tje-TJEJU_2162          -CCCCCHHHHHHHHHHCHHHHCCC-HHHHHHHHHHHHHHHHHCEEEEECCEE-CHHHHHH
28_prn-BW723_13825         -CCCCCHHHHHHHHHHCHHHHCCC-HHHHHHHHHHHHHHHHHCEEEEECCEE-CHHHHHH
30_aqb-D1818_09550         CCCCCCCHHHHHHHHHCHHHHCCC-HHHHHHHHHHHHHHHHHCEEEECCCEECCHHHHHH
31_laci-CW733_06300        -CCCCCHHHHHHHHHHCHHHHCCC-HHHHHHHHHHHHHHHHHCEEEEECCEE-CHHHHHH
36_cep-Cri9333_2611        -CCCCCHHHHHHHHHHCHHHHCCC-HHHHHHHHHHHHHHHHHCEEEECCCEECCHHHHHH
37_ava-Ava_0677            -CCCCCHHHHHHHHHHCHHHCCCC-HHHHHHHHHHHHHHHHHCEEEECCCEECCHHHHHH
40_cthe-Chro_1359          -CCCCCHHHHHHHHHHCHHHCCCC-HHHHHHHHHHHHHHHHHCEEEECCCEECCHHHHHH
41_ptn-PTRA_a1372          -CCCCCHHHHHHHHHHCHHHHCCC-HHHHHHHHHHHHHHHHHCEEEECCCEECCHHHHHH
45_rhh-E0Z06_01765         -CCCCCHHHHHHHHHHCHHHHCCC-HHHHHHHHHHHHHHHHHCEEEECCCEECCHHHHHH
49_nii-Nit79A3_3401        -CCCCCCHHHHHHHHHCHHHHCCC-HHHHHHHHHHHHHHHHHCEEEECCCCCCCHHHHHH
50_ifl-C1H71_04530         -ACCCCCHHHHHHHHHCHHHHCCC-HHHHHHHHHHHHHHHHHCEEEECCCCCCCHHHHHH
53_mpsy-CEK71_12595        --CCCCHHHHHHHHHHCHHHCCCC-HHHHHHHHHHHHHHHHHCEEEECCCCECCHHHHHH
57_dal-Dalk_5292           -CCCCCHHHHHHHHHHCHHHCCCC-HHHHHHHHHHHHHHHHHCEEEECCCCECCHHHHHH
62_pol-Bpro_2870           -CCCCCHHHHHHHHHHCHHHCCCC-HHHHHHHHHHHHHHHHHCEEEECCCCCCCHHHHHH
63_eba-ebA6530             -CCCCCHHHHHHHHHHCHHHCCCC-HHHHHHHHHHHHHHHHHCEEEECCCCECCHHHHHH
64_adi-B5T_02407           -CCCCCHHHHHHHHHHCCCCCCCC-HHHHHHHHHHHHHHHHHCEEEECCCEECCHHHHHH
65_haz-A9404_06910         -CCCCCHHHHHHHHHHCHHHCCCC-HHHHHHHHHHHHHHHHHCEEEECCCCECCHHHHHH
66_bfz-BAU07_18345         -CCCCCHHHHHHHHHHCHHHCCCC-HHHHHHHHHHHHHHHHHCEEEECCCEECCHHHHHH
69_csa-Csal_2711           -CCCCCHHHHHHHHHHCHHHCCCC-HHHHHHHHHHHHHHHHHCEEEECCCEECCHHHHHH
71_pmex-H4W19_15985        -CCCCCHHHHHHHHHHCHHHHCCC-HHHHHHHHHHHHHHHHHCEEEECCCCECCHHHHHH
```

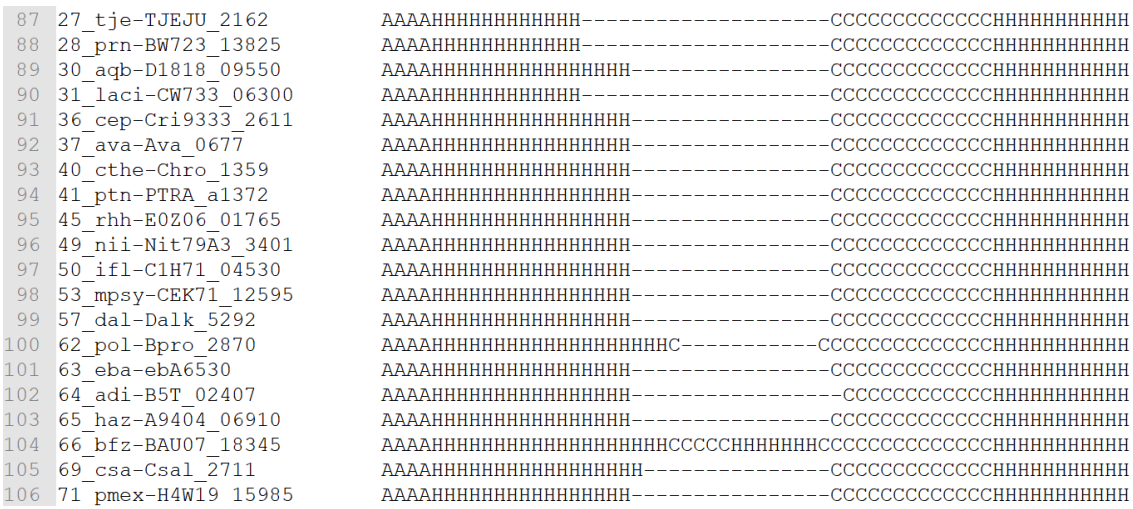
Fig S4. Continuación…

```
27_tje-TJEJU_2162          AAAAHHHHHH----CCCCHHHCCCCEEEECCCCCE---ECCEEECCCEEEEECC--ECCC
28_prn-BW723_13825         AAAAHHHHHH----CCCHHHCCCCCEEEECCCCEE---ECCEEECCCCEEECCCC--CCC
30_aqb-D1818_09550         AAAAHHHHHH----CCCHHHHCCCCEEEECCCCEE---ECCEEEECCCEECCCCCE----
31_laci-CW733_06300        AAAAHHHHHH----CCCHHHHCCCCEEEECCCCEE---ECCEEECCCCEEEECCC--CCC
36_cep-Cri9333_2611        AAAAHHHHHH----CCCHHHHCCCCEEEECCCCEE---ECCEEEECCEEECCCCCCCCCC
37_ava-Ava_0677            AAAAHHHHHH----CCCHHHHCCCCEEEECCCCEE---ECCEEEECCCEECCCCCCECCC
40_cthe-Chro_1359          AAAAHHHHHH----CCCHHHHCCCCEEEECCCCEE---ECCEEECCCCEEECCCCEEECCC
41_ptn-PTRA_a1372          AAAAHHHHHH----CCCHHHHCCCCEEEECCCCEE---ECCEEECCCCEEECCCCEECCC
45_rhh-E0Z06_01765         AAAAHHHHHH----CCCHHHCCCCCEEEECCCCEE---ECCEEEECCCEECCCCCEECCC
49_nii-Nit79A3_3401        AAAAHHHHHH----CCCHHHHCCCCEEEECCCCEE---ECCEEECCCCEEECCCCEECCC
50_ifl-C1H71_04530         AAAAHHHHHH----CCCHHHHCCCCEEEECCCCEE---ECCEEEECCCEECCCCCEECEE
53_mpsy-CEK71_12595        AAAAHHHHHH----CCCHHHHCCCCEEEECCCCEE---ECCEEEECCCEECCCCCCCCCC
57_dal-Dalk_5292           AAAAHHHHHH----CCCHHHCCCCCEEEECCCCCECCCCCCEECCCCEEEEECCEECCC
62_pol-Bpro_2870           AAAAHHHHHH----CCCHHHHCCCCEEEECCCCEE---ECCEEECCCCEEEEECCEECCC
63_eba-ebA6530             AAAAHHHHHH----CCCHHHHCCCCEEEECCCCEE---ECCEEEECCCEECCCCCCECEE
64_adi-B5T_02407           AAAAHHHHHH----CCCHHHHCCCCEEEECCCCEE---ECCEEEECCCEECCCCCCCCCC
65_haz-A9404_06910         AAAAHHHHHH----CCCHHHHCCCCEEEECCCCEE---ECCEEEECCCEECCCCCCCCEE
66_bfz-BAU07_18345         AAAAHHHHHH----CCCHHHHCCCCEEEECCCCEE---ECCEEEECCCEECCCCCCCCCE
69_csa-Csal_2711           AAAAHHHHHHCCCCCCHHHCCCCEEEECCCCEE---ECCEEECCCCEEEECCCEECCC
71_pmex-H4W19_15985        AAAAHHHHHH----CCCHHHCCCCCEEEECCCCEE---ECCEEECCCCEEECCCCCCCCC
```

Continuación…

```
27_tje-TJEJU_2162          CCCCCCEEEEHHHHHHHHCCCCCCCCHHHHHHHHHHHHHHCCCCCCCCCCCCCC-----H
28_prn-BW723_13825         CCCCCCEEEEHHHHHHHHCCCCCCCCEEHHHHHHHHHHHHCCCCCCCCCCCCCHH-----H
30_aqb-D1818_09550         ------EEEEHHHHHHHHHCCCCCCCEEHHHHHHHHHHHHCCCCCCCCCCHC---------
31_laci-CW733_06300        CCCCCEEEEEHHHHHHHHCCCCCCCCHHHHHHHHHHHHHCCCCCCCCCCCCCCH-----H
36_cep-Cri9333_2611        CCCCCEEEEEHHHHHHHHCCCCCCCCEEHHHHHHHHHHHHCCCCCCCCCCCCCCH-----H
37_ava-Ava_0677            CCCCCCEEEEHHHHHHHHCCCCCCCCEEHHHHHHHHHHHHCCCCCCCCCCCCCCH-----H
40_cthe-Chro_1359          CCCCCCEEEEHHHHHHHHCCCCCCCCEEHHHHHHHHHHHHCCCCCCCCCCCCCCC-----H
41_ptn-PTRA_a1372          CCCCCCEEEEHHHHHHHHCCCCCCCEEEEHHHHHHHHHHHCCCCCCCCCCCCCCC-----H
45_rhh-E0Z06_01765         CCCCCCEEEEHHHHHHHHCCCCCCCCEEHHHHHHHHHHHHCCCCCCCCCCCCCCC-----H
49_nii-Nit79A3_3401        CCCCCCEEEEHHHHHH--CCCCCCCCEEHHHHHHHHHHHHCCCCCCCCCCCCCCC-----H
50_ifl-C1H71_04530         CCCCCEEEEEHHHHHHHH--CCCCCCEEEHHHHHHHHHHHCCCCCCCCCCCCCCC-----H
53_mpsy-CEK71_12595        CCCCCCEEEEHHHHHHHHCCCCCCCCEEEHHHHHHHHHHHCCCCCCCCCCCCCCC-----H
57_dal-Dalk_5292           CCCCCCEEEEHHHHHHHHHCCCCCCEEEEHHHHHHHHHHHCCCCCCCCCCCCCCH-----H
62_pol-Bpro_2870           CCCCCCEEEEHHHHHHHHHCCCCCCEEEHHHHHHHHHHHCCCCCCCCCCCCCCC-----H
63_eba-ebA6530             GGCCCEEEEECHHHHH----CCCCCCEEHHHHHHHHHHHCCCCCCCCCCCCCC-----H
64_adi-B5T_02407           CCCCCCEEEEHHHHHHHH--CCCCCCEEEHHHHHHHHHHCCCCCCCCCCCCCC-----H
65_haz-A9404_06910         CCCCCCEEEEHHHHHHHHCCCCCCCCEEEHHHHHHHHHHCCCCCCCCCCCCCC-----H
66_bfz-BAU07_18345         GCCCCCEEEEHHHHH----CCCCCCEEEEHHHHHHHHHHCCCCCCCCCCCCCCCCHHHH
69_csa-Csal_2711           CCCCCEEEEEHHHHHHH--CCCCCCCEEHHHHHHHHHHCCCCCCCCCCCCCC-----H
71_pmex-H4W19_15985        CCCCCCEEEEHHHHHHHHCCCCCCCCEE-HHHHHHHHHHCCCCCCCCCCCCHHH-----H
```

Fig S4. Continuación…



Continuación…

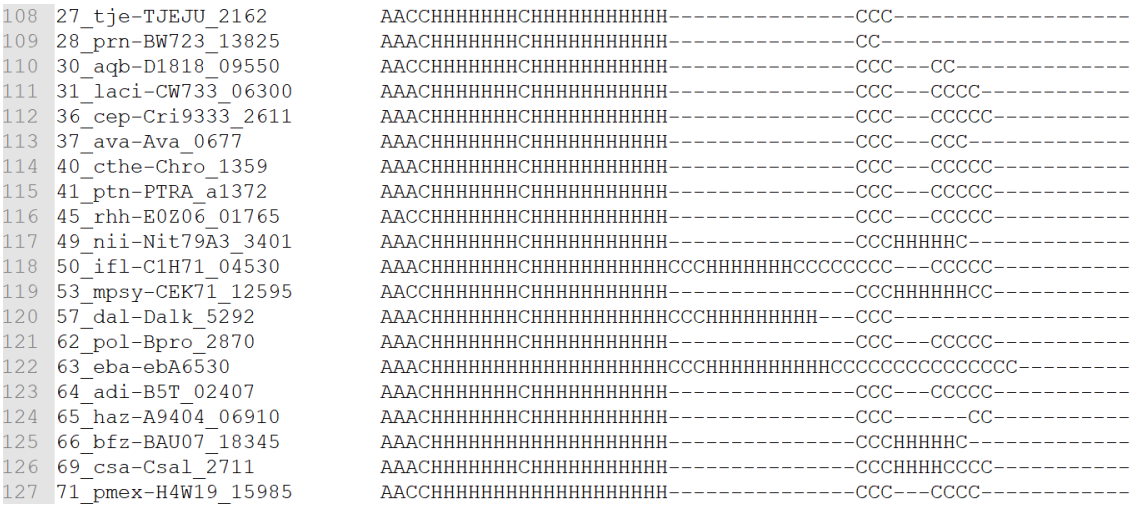


Fig S4. Continuación…

```
129 27_tje-TJEJU_2162         --------
130 28_prn-BW723_13825        --------
131 30_aqb-D1818_09550        --------
132 31_laci-CW733_06300       --------
133 36_cep-Cri9333_2611       --------
134 37_ava-Ava_0677           --------
135 40_cthe-Chro_1359         --------
136 41_ptn-PTRA_a1372         --------
137 45_rhh-E0Z06_01765        --------
138 49_nii-Nit79A3_3401       --------
139 50_ifl-C1H71_04530        --------
140 53_mpsy-CEK71_12595       --------
141 57_dal-Dalk_5292          --------
142 62_pol-Bpro_2870          --------
143 63_eba-ebA6530            --------
144 64_adi-B5T_02407          --------
145 65_haz-A9404_06910        --------
146 66_bfz-BAU07_18345        --------
147 69_csa-Csal_2711          --------
148 71_pmex-H4W19_15985       --------
```

C-terminal

Fig S4. Alineamiento múltiple de los elementos de estructura secundaria de las proteínas del COG3228 con respecto al contenido de GC de sus genes. El formato de alineamiento se divide en dos partes: la primera parte de lado izquierdo refleja el valor (orden ascendente) del contenido de GC de los genes que codifican a las proteínas del COG, seguido del nombre del gen; la segunda parte muestra el alineamiento de las secuencias de estructuras secundarias: H para hélice alfa, E para hoja beta, y C para lazo.

Anexo

Anexo 1. Selección de especies no redundantes. Las cepas de *Candidatus Sulcia muelleri* existentes en la base de datos KEGG. El ID de KEGG (columna uno), las cepas (columna 2), el tamaño del genoma (bp) (columna 3), el número de proteínas (columna 4) y el contenido de GC genómico (columna 5).

ID_KEGG	cepas	Tamaño del genoma	# proteinas	%GC genómico
sum	Candidatus Sulcia muelleri CARI	276511	246	20.95
smup	Candidatus Sulcia muelleri PSPU	285352	251	20.83
smum	Candidatus Sulcia muelleri ML	190405	187	23.51
smv	Candidatus Sulcia muelleri Sulcia-ALF	190733	188	23.39
smue	Candidatus Sulcia muelleri TETUND	270029	247	22.79
sms	Candidatus Sulcia muelleri SMDSEM	276984	242	22.49
smub	Candidatus Sulcia muelleri BGSS	244618	227	22.24
smg	Candidatus Sulcia muelleri GWSS	245530	227	22.17
smh	Candidatus Sulcia muelleri DMIN	243933	226	22.11

Dos cepas de Candidatus *Sulcia muelleri* fueron elegidas (sum y smup) por tener tamaños de genomas grandes (respectivamente), mayor número de proteínas y porque contienen el mismo contenido de GC genómico (de 20%).

Reconocimiento especial al apoyo brindado para la realización de esta tesis:

A la M.B. María Luisa Tabche Barrera, por su apoyo en el buen funcionamiento del laboratorio y su organización para contar de manera adecuada todos los reactivos utilizados durante una primera parte experimental de mis estudios doctorales. Por cuestiones de las restricciones de seguridad sanitaria impuestas para contender con la pandemia causada por el virus SARS-CoV-2, se decidió realizar un proyecto teórico que es el que constituye la tesis que aquí se presenta.

Al M. en C. Walter Josué Hernández Santos por su apoyo en la realización de la página web Gcto2D (https://biocomputo.ibt.unam.mx/gcto2d/).

Al M. en C. Fernando Fontove Herrera por su asesoría en el desarrollo de algoritmos computacionales y análisis estadísticos.

12. BIBLIOGRAFÍA

1. Sueoka N. Correlation between Base Composition of Deoxyribonucleic Acid and Amino Acid Composition of Protein. *PROC N A S*. 1961;47:1141-1129.
2. Chou PY, Fasman GD. Conformational parameters for amino acids in helical, Beta-sheet, and random coil region calculated from proteins. *Biochemetry*. 1974;13(2):211-222.
3. Chou PY, Fasman GD. Prediction of Protein Conformation. *Biochemistry*. 1974;13(2):222-245. doi:10.1021/bi00699a002
4. Wu H, Zhang Z, Hu S, Yu J. On the molecular mechanism of GC content variation among eubacterial genomes. *Biol Direct*. 2012;7(2):1-16. doi:10.1186/1745-6150-7-2
5. McCutcheon JP, Moran NA. Functional convergence in reduced genomes of bacterial symbionts spanning 200 my of evolution. *Genome Biol Evol*. 2010;2(1):708-718. doi:10.1093/gbe/evq055
6. Thomas SH, Wagner RD, Arakaki AK, et al. The Mosaic genome Anaeromyxobacter dehalogenans 2CP-C suggest an aerobic common ancestor to the Delta-Proteobacteria. *PLoS One*. 2008;3(5):e2103.
7. Musto H, Naya H, Zavala A, Romero H, Alvarez-Valín F, Bernardi G. Genomic GC level, optimal growth temperature, and genome size in prokaryotes. *Biochem Biophys Res Commun*. 2006;347(1):1-3. doi:10.1016/j.bbrc.2006.06.054
8. Almpanis A, Swain M, Gatherer D, McEwan N. Correlation between bacterial G+C content, genome size and the G+C content of associated plasmids and bacteriophages. *Microb genomics*. 2018;4(4):e000168. doi:10.1099/mgen.0.000168
9. McCutcheon JP, McDonald BR, Moran NA. Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. *PLoS Genet*. 2009;5(7). doi:10.1371/journal.pgen.1000565
10. Mann S, Chen YPP. Bacterial genomic G + C composition-eliciting enviromental adaptation. *Genomics*. 2010;95:7-15. doi:10.1371/journal.pone.0107319
11. Foerstner KU, von Mering C, Hooper SD, Bork P. Environments shape the

nucleotide composition of genomes. *EMBO Rep*. 2005;6(12):1208-1213. doi:10.1038/sj.embor.7400538

12. Chen W, Shao Y, Chen F. Evolution of complete proteomes: Guanine-cytosine pressure, phylogeny and environmental influences blend the proteomic architecture. *BMC Evol Biol*. 2013;13(1):1. doi:10.1186/1471-2148-13-219
13. Zhou HQ, Ning LW, Zhang HX, Guo FB. Analysis of the relationship between genomic GC content and patterns of base usage, codon usage and amino acid usage in prokaryotes: Similar GC content adopts similar compositional frequencies regardless of the phylogenetic lineages. *PLoS One*. 2014;9(9). doi:10.1371/journal.pone.0107319
14. Galtier N, Lobry JR. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *J Mol Evol*. 1997;44(6):632-636. doi:10.1007/PL00006186
15. Kagawa Y, Nojima H, Nukiwa N, et al. High guanine plus cytosine content in the third letter of codons of an extreme thermophile. DNA sequence of the isopropylmalate dehydrogenase of Thermus thermophilus. *J Biol Chem*. 1984;259(5):2956-2960. doi:10.1016/s0021-9258(17)43242-x
16. Lightfield J, Fram NR, Ely B. Across bacterial phyla, distantly-related genomes with similar genomic GC content have similar patterns of amino acid usage. *PLoS One*. 2011;6(3). doi:10.1371/journal.pone.0017677
17. Muto A, Osawa S. The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc Natl Acad Sci U S A*. 1987;84(1):166-169. doi:10.1073/pnas.84.1.166
18. Sueoka N. Directional mutation pressure and neutral molecular evolution. *Proc Natl Acad Sci U S A*. 1988;85(8):2653-2657. doi:10.1073/pnas.85.8.2653
19. Gu X, Hewett-emmett D, Li W. Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. *Genetica*. 1998;102/103:383-391.
20. Singer GAC, Hickey DA. Nucleotide Bias Causes a Genomewide Bias in the Amino Acid Composition of Proteins. *Mol Biol Evol*. 2000;17(11):1581-1588.
21. Andersson SGE, Sharp PM. Codon usage and base composition in Rickettsia

prowazekii. *J Mol Evol*. 1996;42(5):525-536. doi:10.1007/BF02352282

22. Lobry JR. Influence of genomic G + C content on average amino-acid composition of proteins from 59 bacterial species. *Gene*. 1997;205(1-2):309-316. doi:10.1016/S0378-1119(97)00403-4
23. Bernardi G, Bernardi G. Compositional Constraints and Genome Evolution *. *J Mol Evol*. 1986;24:1-11.
24. Chou PY, Fasman GD. Prediction of the Secondary Structure of Proteins From Their Amino Acid Sequence. *Adv Enzymol Relat Areas Mol Biol*. 1978;47(1195):45-148. doi:10.1002/9780470122921.ch2
25. Pirovano W, Heringa J. *Protein Secondary Structure Prediction*. Vol 609.; 2010. doi:10.2307/j.ctvfxvccj.23
26. Argos P, Palau J. Amino acid distribution in protein secondary structures. *Int J Pept Protein Res*. 1982;19(4):380-393. doi:10.1111/j.1399-3011.1982.tb02619.x
27. Lagerkvist U. “Two out of three”: An alternative method for codon reading. *Proc Natl Acad Sci U S A*. 1978;75(4):1759-1762. doi:10.1073/pnas.75.4.1759
28. Lehmann J, Libchaber A. Degeneracy of the genetic code and stability of the base pair at the second position of the anticodon. *Rna*. 2008;14(7):1264-1269. doi:10.1261/rna.1029808
29. Bernardi G, Bernardi G. Codon Usage and Genome Composition. *J Mol Evol*. 1985;22:363-365.
30. Codon Usage Database. https://www.kazusa.or.jp/codon/. Accessed February 9, 2023.
31. Knight RD, Freeland SJ, Landweber LF. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol*. 2001;2(4):1-13. doi:10.1186/gb-2001-2-4-research0010
32. McCutcheon JP. The bacterial essence of tiny symbiont genomes. *Curr Opin Microbiol*. 2010;13(1):73-78. doi:10.1016/j.mib.2009.12.002
33. Bennett GM, Moran NA. Small, smaller, smallest: The origins and evolution of ancient dual symbioses in a phloem-feeding insect. *Genome Biol Evol*. 2013;5(9):1675-1688.

34. Garcia Costas AM, Liu Z, Tomsho LP, Schuster SC, Ward DM, Bryant DA. Complete genome of Candidatus Chloracidobacterium thermophilum, a chlorophyll-based photoheterotroph belonging to the phylum Acidobacteria. *Environ Microbiol*. 2012;14(1):177-190. doi:10.1111/j.1462-2920.2011.02592.x
35. Zhao X, Zhang Z, Yan J, Yu J. GC content variability of eubacteria is governed by the pol III α subunit. *Biochem Biophys Res Commun*. 2007;356(1):20-25. doi:10.1016/j.bbrc.2007.02.109
36. Oliver JL, Marín A. A relationship between GC content and coding-sequence length. *J Mol Evol*. 1996;43(3):216-223. doi:10.1007/BF02338829
37. Naya H, Romero H, Zavala A, Alvarez B, Musto H. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. *J Mol Evol*. 2002;55(3):260-264. doi:10.1007/s00239-002-2323-3
38. Sueoka N. On the Genetic basis of Variation and heterogeneity of DNA bases of Composition. *PROC N A S*. 1962;48:582-592.
39. Raghavan R, Kelkar YD, Ochman H. A selective force favoring increased G+C content in bacterial genes. *Proc Natl Acad Sci U S A*. 2012;109(36):14504-14507. doi:10.1073/pnas.1205683109
40. Freese E. On the evolution of the base composition of DNA. *J Theor Biol*. 1962;3(1):82-101. doi:10.1016/S0022-5193(62)80005-8
41. Singer CE, Ames BN. Sunlight Ultraviolet and Bacterial DNA Base Ratios We have found a strong correlation. *Science (80-)*. 1970;170(3960):822-826.
42. McEwan CEA, Gatherer D, McEwan NR. Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. *Hereditas*. 1998;128(2):173-178. doi:10.1111/j.1601-5223.1998.00173.x
43. Madigan MT, Martinko JM, Dumlap P V., Clark DP. *Brock: Biología de Los Microorganismos*. 12th ed. Pearson; 2009.
44. Alberts B, Johnson A, Lewis J, et al. *Molecular Biology of the Cell*. 6th ed. Garlan Science; 2014.
45. Lewis PN, Go N, Go M, Kotelchuck D, Scheraga HA. Helix probability profiles of denatured proteins and their correlation with native structures. *Proc Natl Acad Sci U S A*. 1970;65(4):810-815. doi:10.1073/pnas.65.4.810

46. ACS publications Most trusted, Most Cited and Most Read. https://pubs.acs.org/doi/10.1021/bi00699a001#. Accessed January 21, 2023.
47. KEGG: Kyoto Encyclopedia of Genes and Genomes. https://www.genome.jp/kegg/. Accessed June 6, 2022.
48. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/. Accessed June 10, 2022.
49. UniProt. https://www.uniprot.org/. Accessed October 5, 2021.
50. Yan R, Xu D, Walker S, Zhang Y. PSSpred · bio.tools. https://bio.tools/psspred. Accessed October 5, 2021.
51. PSIPRED Workbench. http://bioinf.cs.ucl.ac.uk/psipred/. Accessed June 10, 2022.
52. JPred: A Protein Secondary Structure Prediction Server. https://www.compbio.dundee.ac.uk/jpred/. Accessed June 10, 2022.
53. SWISS-MODEL Interactive Workspace. https://swissmodel.expasy.org/interactive. Accessed June 10, 2022.
54. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 2000;28(1):27-30. https://www.genome.jp/kegg/. Accessed June 10, 2022.
55. Yan R, Xu D, Yang J, Walker S, Zhang Y. A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. *Sci Rep*. 2013;3:2619. doi:10.1038/srep02619
56. Tatusov RL, Fedorova ND, Jackson JD, et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*. 2003;4(4):1-14.
57. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32(5):1792-1797. doi:10.1093/nar/gkh340
58. Eddy SR. Profile hidden Markov models. *Bioinformatics*. 1998;14(9):755-763. doi:10.1093/bioinformatics/14.9.755
59. McKinney W. Data Structures for Statistical Computing in Python. *Proc 9th Python Sci Conf*. 2010;1(Scipy):56-61. doi:10.25080/majora-92bf1922-00a
60. Harris CR, Millman KJ, van der Walt SJ, et al. Array programming with NumPy. *Nature*. 2020;585(7825):357-362. doi:10.1038/s41586-020-2649-2
61. Levitt M. Conformational Preferences of Amino Acids in Globular Proteins.

Biochemistry. 1978;17(20):4277-4285. doi:10.1021/bi00613a026

62. Mira A, Martín-Cuadrado AB, D'Auria G, Rodríguez-Valera F. The bacterial pan-genome: A new paradigm in microbiology. *Int Microbiol*. 2010;13(2):45-57. doi:10.2436/20.1501.01.110

63. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589. doi:10.1038/s41586-021-03819-2

DRA. LINA ANDREA RIVILLAS ACEVEDO
COORDINADORA DEL POSGRADO EN CIENCIAS
PRESENTE

Atendiendo a la solicitud para emitir DICTAMEN sobre la revisión de la tesis titulada: Efecto del sesgo del contenido de GC genómico de organismos procariotas en las estructuras secundarias de sus proteínas, que presenta la alumna Diana Barceló Antemate **(**10010088**)** para obtener el título de **Doctor en Ciencias.**

Director de tesis: Dr. Enrique Merino Pérez

Nos permitimos informarle que nuestro voto es:

NOMBRE	DICTAMEN	FIRMA
Dra. Carmen Nina Pastor Colón CIDC – UAEM	APROBADO	13/abril/2023
Dra. María del Rayo Sánchez Carbente CEIB – UAEM	APROBADO	27/marzo/2023
Dr. Ramón Antonio Gonzalez García Conde CIDC – UAEM	APROBADO	23/marzo/2023
Dra. Cinthia Ernestina Nuñez López IBT – UNAM	APROBADO	24/marzo/2023
Dra. Verónica Mercedes Narváez Padilla CIDC – UAEM	APROBADO	17/abril/2023
Dra. Rosa María Gutiérrez Ríos IBT – UNAM	APROBADO	10/abril/2023
Dr. Armando Hernández Mendoza CIDC – UAEM	APROBADO	24/marzo/2023

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

RAMON ANTONIO GONZALEZ GARCIA CONDE | Fecha:2023-03-23 18:40:39 | Firmante

LtrHL576vB82eiNwHxNELH7InfxLp90odeuLJYpdQ2C2Le7YSTTI6baZinzn0eUUE+zcrcuKWnJB9K8DUaSgRpU/or1qqkxHdgklBltxjn5aQHA7q4IsVmvrkUIZMcu1/57kvVlu5K2aA0GQlcaXMkN55+NIBCnrZkPoh1/TThTqFYuK3TIUNI6ComDcyMNK3OE7arDYjnvA0QeDtvcqf6u+BtznYSpra98uaf3w/mLp9TeRopcLC2WzlNyLlQR/LkBuyUpEGpW5uIRusxdR7f1wFgaQWrkeJA7Q344fJZOLfPOOc+9wGIe9d3rEZBYX+xQ9Svw4VKWty7//GHtHug==

ARMANDO HERNANDEZ MENDOZA | Fecha:2023-03-24 08:27:59 | Firmante

guUKznBaec2GGKw8XjMCdDzxZuRM6mJgJfz3DfINooeV/c6C184KVnenc194XoutKW86y8ffwYmDENB+Mt26+n14ygpzW0RdKXYk2gC26l/4gDvj6fVjAHvkximz6ss3BLwpCwq8dHE1OZIkZJN9mF1Ldpe2qrkR1seDJ54Hok2HnwGKhs8uM7UhfotoOI70myDDZNue70/oGGJntxbINj4dJJa1UOMVY1a0abEU9Ehnr8qRUFUpJz20SIpkWtiR9S/bzxabgIUcMuplb+VHclGVbD2DAZ4aGo/zDVXv83cxzYCGD9fNr95kIJ2sjsr0579AxMDIYGgGAMrbPuUURw==

CINTHIA ERNESTINA NUÑEZ LOPEZ | Fecha:2023-03-24 08:55:51 | Firmante

kHHTBaEsANISKoASbRFoKdnkztj4heNNOnRb7lz6I7Nzm6Elvsu80F6WkdeA6//SyYx5iN4d9hnXCmvdkJh1visa9859HS9MMEcy/qc46oTM5tNibUnhc8dlRHZP8eIuO9wvZrCWG3LliR52kDG8sCZ5vYl9s7lNFO+iCca9Lu9qVWw9fDxdBZDvVPddKnbTSOqqyn/nFGqV7ZdGErPvaqEXOfUzP+cJno30s0E40sZtS3ElTJZgZkFTI/c7FHFsbMBODDlJjEJsqgLWgtNCQd/+k332uLXB6qugRsIQEDwyj45vaOvPTjgWSCRH0s9yMNLyfg6wit2y+04UWyNtlg==

MARIA DEL RAYO SANCHEZ CARBENTE | Fecha:2023-03-27 16:26:23 | Firmante

Dpn/w3Dv1be/0aFby3PATоHqOJWpU4/NuvABABzC3JjECMnQHxiC9uT5QDUlxKPKrWuatviMay4quQmCsAS5ReFbWCkDEwU50SCJlaQUR2ta3dyaI9v5iQrFPNJ2d4TGlx99NIBUT1xcDOh57mpcoteM2+j0yLM+MRS4qHSJG0LWSraI/saXUG9RJsXPhW7J5oNXgJQAMtbm517/tH6BwOpBzGxKFSzVB+2zwrLMz3Mh/pggjDve61+XwBfOSYaUFHmo+nnjQI7pQhpzL8XzNCEtECF9U33nnMVsyID2EeXkpY0BnH9cl/GQzQnEIy4Btg+JDXAECCB/baWV7HBy5w==

ROSA MARÍA GUTIÉRREZ RÍOS | Fecha:2023-04-10 19:12:44 | Firmante

ZwaMzsO/WR1XJVUHf/I8FvEZsGZ34FgFoWXxXnlHjdFCPoQqL58blyz8XY+OhLKvUamGi6e30+8GB/kZqEBAEYYdbobuku2Asw80H9btf3bvDSRUhxeLTsE6tkOZsBACpAUjtcG94eO/s/E4pZaYtGmvTDRs4BCP8ZfadrqyWR0ammEBjuBYyvSX5k4WGBPCQm+8opQ7Afd8hWmgNl558jzS77G5ofBuWhlVLHuUnzXXYN9E39+wCjfazZ3ABv82pKarv9d3MTv8lygle15ErJjiTuOsSgZeyh9S61JCMlY1PplAYTMm0D6Rfi3PphKHtWaKOz9Azb7MVM9lLIkXUg==

CARMEN NINA PASTOR COLON | Fecha:2023-04-13 17:46:44 | Firmante

SIx8P62o+LvkN7IZ6W+oH4B9KLXaGQ1s6WpR0wQBrDfaf5C2LGPopZvL60ztvVXP6S5UyflfooMRL5uk6584+GqJYahJxMPYhh2SmakhZFE1n6ZyyXNqqd4w/jR0fZbVYZJSSjhCrR8aR1buitVLDoPLO+qEwt7hyKkWys+mbLbEKlLKgAEWTypXMd7rjw22Alzn0jwGlWmeqxLQmAZtHheYtdO/TSiYMYjUypj/9aElUu3z8m2VigpUE04+sUo4ytY6eqpajoHX0MWTI3xf0Krhfxkz0tZxsKSilgICtDx8CgQmObt2oM73T9MsWsCa+NYwHabnJ2buzSpnEY/Nrw==

VERONICA MERCEDES NARVAEZ PADILLA | Fecha:2023-04-17 12:01:16 | Firmante

qXzgSQ9dj71e8TEPCapIFyxXhiBnnvAAUFC+MhiTSj7HcO5yHaC0G3HhFiUn3vs7x+34+zgpLe46TU8drpGKdUaCqeVgvwNY1sWsWmVPvwoqE6mDpenBb6vrM40VQRRJfYbzURMlXtPWqEjU6XUXKPhzMR/diCnPVg1kqD+NVtyIDGLsJp+OwO5fhjUq8CsE+/PWQCqM671kek54sJBz0JSHWI8mO2l1yVarO8dlTJbA8IzqaJU1MtpgtRMtHAw5N6biFfNAuNd7gYcepKZAgxs6xiEqCdCRmYgGgkyI1g9vAGVvG0DcKvV947wW1OeNQkCdqbbqcLfe6Ypl0AS3TA==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

TJ8wk9Mgv

https://efirma.uaem.mx/noRepudio/3ATzRdyCnkrwIBmWQAv94HbHFqg6RIg3