Revista Latinoamericana de Patología Clínica
MEDICINA DE LABORATORIO

Análisis genómico de las funciones proteicas y relaciones taxonómicas de bacterias causantes de enfermedad diarreica aguda

Jeel Jr. Moya-Salazar,* Roberto A Ubidia-Incio*

Palabras clave: Genómica comparativa, COGs, TaxPlot, enfermedad diarreica aguda, bioinformática.

Key words: *Comparative genomics, COGs, TaxPlot, acute diarrheal disease, bioinformatics.*

RESUMEN

Se diseñó un estudio analítico-correlacional de corte transversal para analizar los genomas de bacterias causantes de la enfermedad diarrea aguda (EDAs) y se establecieron relaciones funcionales entre funciones proteicas con base en el análisis realizado con las herramientas COGs data base (http://www.ncbi.nlm.nih.gov/COG/) y TaxPlot (http://www.ncbi.nlm. nih.gov/sutils/taxik2.cgi?isbact=1). Se realizó el conteo de las categorías de los COGs para cada especie, la determinación del porcentaje real de cada categoría respecto a la cantidad total de COGs, un análisis comparativo interproteico de cada especie mediante TaxPlot y la clasificación filogenética de las proteínas codificadas en genomas completos utilizando el algoritmo UPGMA. Se incluyó un total de 15 microorganismos (9 bacterias causantes de EDAs y 6 bacterias saprofitas). 11 de las 15 bacterias presentaron mayor COGs en las categorías J –12.25%– (6 bacterias causantes de EDAs y 5 bacterias saprófitas). Las categorías funcionales para metabolismo fueron uniformes y las categorías funcionales Y (0.00%), Z (0.03%), B (0.01%) y A (0.05%) fueron las de menor cuantía. Los COG pueden extrapolarse en datos exactos que describen el perfil proteico de los agentes patógenos. El método filogenético por TaxPlot reduce el tiempo de análisis y representa una aproximación holística al estudio de las relaciones filogenéticas a nivel proteómico.

ABSTRACT

An analytical-correlational cross section study was designed to analyze the genomes of acute diarrheal disease (ADD) causing bacteria and functional relations were established between protein functions based on the analysis performed by COGs database and TaxPlot tools. We counted the percentages of COGs for every category for each species, and determined the percentages based on the total number of COGs for each species. We made an interproteic comparative analysis for each species using TaxPlot and a phylogenetic classification for the proteins codified in complete genomes using the UPGMA algorithm. A total of 15 microorganisms were included (9 AAD-causing bacteria and 6 saprophytic bacteria). Eleven of these 15 bacteria had higher COGs in the categories J –12.25%– (6 ADD-causing bacteria and 5 saprophytic bacteria). The metabolic functional categories were uniform, and the functional categories Y (0.00%), Z (0.03%), B (0.01%) and A (0.05%) had the smallest quantities. COGs can be extrapolated into exact data that describe the protein profile of these pathogenic agents. The TaxPlot phylogenetic method reduces analysis time and represents an holistic approximation for the study of the phylogenetic relations at a proteomic level.

* Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Lima, Perú.

Correspondencia:
Jeel Jr. Moya-Salazar
Jr. Pacífico 957, Urb. San Felipe,
Lima 07, Lima, Perú.
Tel: 01 4873681
Móvil: 051 1986014954
E-mail: jeel.moya.s@upch.pe

Recibido: 07/08/2015
Aceptado: 17/08/2015

INTRODUCCIÓN

La enfermedad diarrea aguda (EDAs o ADD por sus siglas en inglés) es una de las 4 principales causas de mortalidad infantil y una de las 3 que tienen relación con el cambio climático.[1] En la región de las Américas la mortalidad anual por enfermedad diarreica se estima en 112,000 casos.[2] Esta afección ocasiona alteraciones en el desarrollo humano en 3 aspectos fundamentales: limitaciones en el crecimiento orgánico, disminución del capital intelectual e incremento de los años de vida saludable perdidos (AVISA o DALY por sus siglas en inglés), principalmente en niños menores de 2 años.[3]

Las principales causas de EDAs son bacterias del género *Campylobacter*, *Salmonella*, *Shigella*, *Vibrio*, *Proteus* y *Enterococcus faecalis*; además de *Escherichia coli* O157:H7, *Klebsiella pneumoniae*, *Morganella morganii*, *Aeromona oxytona*, entre otras.[2,4-7]

El análisis de genomas resulta útil porque permite entender los diversos fenómenos biológicos al mostrar la conservación de genes responsables de funciones específicas, no de un solo gen aislado sino como un grupo con una función que se ha conservado como unidad a lo largo de la evolución.[8]

El tener genomas de varios organismos relacionados permite también observar y comprender cómo la aparición de nuevos genes puede llevar a desarrollar nuevas funciones dentro de un sistema de genes preestablecido, de esta manera la identificación de ortólogos es crucial para la predicción de funciones en nuevas cepas o especies que causen una patología determinada.[8] Sin lugar a duda, esto último beneficiaría el entendimiento de los mecanismos de patogenicidad utilizados por diversos organismos, como las bacterias causantes de EDAs, la forma en que han aparecido y se han diversificado y en mayor medida, la utilidad al momento de buscar estrategias para combatirlas y prevenir su infección.

La genómica comparativa nos brinda un alcance muy útil para entender la información a nivel genómico, por ello el primer objetivo cuando se tiene un genoma secuenciado es identificar las regiones funcionales, tanto genes como secuencias regulatorias. Parte de esa información requiere un trabajo experimental y parte puede obtenerse a partir de análisis comparativos *in silico* que facilitan la identificación de estos elementos. De esta manera, el propósito de la genómica comparativa es, a través de la comparación de genomas de especies diferentes, entender cómo las especies han evolucionado y qué funciones cumplen los genes y regiones no codantes.[9]

La base de datos *Clusters of Orthologous Groups of proteins* (COG *data base*) y TaxPlot son herramientas bioinformáticas que permiten observar la estructuración del genoma y la similitud que posee un microorganismo con otros respectivamente de acuerdo con secuencias completas[10] y de manera adicional, comparar la similitud entre proteínas del genoma entre diferentes especies. Estas herramientas nos permiten realizar el análisis de genomas en los que previamente se ha hecho un trabajo de clasificación (COGs) o de comparación (TaxPlot), por lo que no existe la necesidad de descargar genomas completos para realizar un análisis; sin embargo, presenta como desventaja el hecho de no tener un mantenimiento continuo, por lo que el número de genomas disponibles para análisis es limitado y desactualizado.[10]

COGs permite observar la distribución y proporción de los genes de un organismo con base en categorías funcionales que son agrupaciones basadas en los mecanismos en que están involucrados los productos de esos genes (relacionadas con el metabolismo, estructura, procesos celulares o manejo de información); estas categorías son clasificadas con letras mayúsculas y pueden estar relacionadas con los mecanismos de expresión genética (J, A, K, L, B), con las funciones celulares no metabólicas (D, Y, V, T, M, N, Z, W, U, O, X) y/o con el metabolismo (C, G, E, F, H, I, P, Q), como se describe en el *cuadro I*.[11] Además, permite observar relaciones con otras especies que contienen genes ortólogos a los de nuestra(s) especie(s) de interés que están agrupados en cada COG (COG 0001-COG 5665).[11,12]

Del mismo modo, la herramienta TaxPlot del *National Center for Biotechnology Information* (NCBI) refiere la cantidad de proteínas similares a partir de los genomas de tres especies utilizando una primera como referencia (*query*) de la cual se toman todas las secuencias de proteínas y se comparan con resultados precomputarizados de BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) con las proteínas de las otras dos especies.[13] Este tipo de análisis se ha utilizado anteriormente para estudiar similitudes entre otros organismos como *Archaeas* [14] y en otros casos para predecir genes de genosomas recién secuenciados como *Serratia plymuthica* AS 12[15] *Staphylococcus epidermidis*,[9] entre otros.[16-18]

El objetivo de la presente investigación fue analizar los genomas de las bacterias causantes de EDAs y establecer relaciones funcionales entre genes y funciones proteicas con base en el análisis realizado con COGs *data base* y TaxPlot.

MATERIAL Y MÉTODOS

Tipo y diseño de investigación: estudio analítico-correlacional, retrospectivo de corte transversal.

Obtención y análisis del genoma de COGs *data base*: se obtuvieron las listas de COGs completas de 9 especies que provocan EDAs desde la base de datos del NCBI (http://www.ncbi.nlm.nih.gov/COG/) y fueron codificadas y tabuladas en MS-Excel 2010 para Windows, las especies consideradas en el presente estudio fueron: *Campylobacter jejuni* subsp. *jejuni* (NCTC 11168 = ATCC 700819) **[Camjej]**, *Shigella dysenteriae* (cepa 1617) **[Shidys]**, *Salmonella enterica* subsp. *enterica* (serovar *typhimurium* cepa LT2) **[Salent]**, *Vibrio cholerae* (01 biovar El Tor cepa N16961) **[Vibcho]**, *Escherichia coli* O157:H7 (cepa Sakai) **[Escc01]**, *Proteus mirabilis* (cepa H14320) **[Promir]**, *Enterococcus faecalis* (cepa V583) **[Entfae]**, *Klebsiella pneumoniae* (cepa 342) **[Klepne]** y *Morganella morganii* subsp. *morganii* (cepa KT) **[Mormor]**.[4,7]

Especies control: como control para el análisis de COGs se utilizaron 6 especies de bacterias no causantes de EDAs: *Haemophilus influenzae* (cepa Rd Kw20) **[Haeinf]**, *Escherichia coli* (cepa K-12, subcepa MG1655, no patógena) **[Esccol]**, *Veillonella parvula* (cepa DSM 2008) **[Veipar]**,

Lactobacillus plantarum (cepa ZJ316) **[Lacpla]**, *Fusobacterium nucleatum* (subsp. *nucleatum* ATCC 25586) **[Fusnuc]** y *Porphyromonas gingivalis* (cepa TDC60) **[Porgin]**.[19]

Recursos para el análisis del genoma

COG *data base*: se realizó el conteo de las categorías de los COGs (http://www.ncbi.nlm.nih.gov/COG/) con ayuda de tablas dinámicas, se generaron dos tablas, una para cada categoría separada en cada especie mostrando la cantidad de COGs para cada especie por categoría, contando repetidamente aquellos COGs con más de una categoría; y el porcentaje real de cada categoría respecto a la cantidad total de COGs para cada especie *(cuadro I)*; y otra para las bacterias saprófitas utilizadas como control.

Cuadro I. Lista de las categorías de la base de datos de COGs.

Almacenamiento y procesamiento de información	
J	Traducción, estructura ribosomal y biogénesis
A	Procesamiento y modificación de ARN
K	Transcripción
L	Replicación, recombinación y reparación
B	Estructura y dinámica de cromatina
Procesos celulares y señalización	
D	Control del ciclo celular, división, partición cromosómica
Y	Estructura nuclear
V	Mecanismos de defensa
T	Mecanismos de transducción de señal
M	Biogénesis de pared celular, membranas y envolturas
N	Motilidad celular
Z	Citoesqueleto
W	Estructuras extracelulares
U	Tráfico celular, secreción y transporte vesicular
O	Modificaciones postraduccionales, recambio de proteínas y chaperonas
X	Mobiloma: profagos y transposones
Metabolismo	
C	Producción y conversión de energía
G	Transporte y metabolismo de carbohidratos
E	Transporte y metabolismo de aminoácidos
F	Transporte y metabolismo de nucleótidos
H	Transporte y metabolismo de coenzimas
I	Transporte y metabolismo de lípidos
P	Transporte y metabolismo de iones inorgánicos
Q	Biosíntesis, transporte y catabolismo de metabolitos secundarios
Poco caracterizados	
R	Sólo predicción de funciones generales
S	Función desconocida

TaxPlot: se realizó un análisis comparativo interproteico de cada especie mediante la herramienta TaxPlot del NCBI (http://www.ncbi.nlm.nih.gov/sutils/taxik2.cgi?isbact=1) con el objetivo de verificar el grado de similitud entre las proteínas ortólogas de las especies analizadas. Se observó la relación de cada especie estudiada (especie «*query*») con las otras en grupos de dos (especies de comparación). Se utilizaron las mismas cepas para todas las especies, excepto *Shigella dysenteriae*, en cuyo caso se encontró sólo la cepa S197 y *Morganella morganii* (cepa KT) que no estaba disponible y fue descartada para el análisis.

Técnica de recolección de datos y procesamiento

Codificación y tabulación de datos

Todos los datos fueron tabulados en la matriz de codificación inicial en MS-Excel 2010, en la que se codificaron con las variables: [Camjej], [Shidys], [Salent], [Vibcho], [Escc01], [Promir], [Entfae], [Klepne] y [Mormor]. Con los datos de COGs se realizó la valoración y la porcentualización de cada grupo funcional para cada microorganismo mediante tablas dinámicas y filtros para los valores de filas y columnas, con los cuales se generaron dos tablas.

Se compararon valores promedio (cantidades reales y porcentajes) entre las bacterias causantes y no causantes de EDAs y además entre los dos grupos más grandes presentes en ambos grupos (*Enterobacteriaceae* y *Firmicutes*) con el fin de tener un mejor panorama analítico y poder discernir las relaciones entre *clados* y la relación por patogenicidad.

Para los datos de TaxPlot se utilizó el número de aciertos (*hits*) totales, esto es el número de proteínas que produjeron aciertos para las dos especies a partir del número total de proteínas de la primera sin tomar en cuenta el sesgo en la similitud hacia alguna de las especies comparadas; todos éstos fueron tabulados en la matriz de codificación inicial contrastados con las variables [Cje], [Sdy], [Sen], [Vch], [Eco], [Pmi], [Efa] y [Kpn], en las que se analizaron las relaciones entre el grado de similitud encontrado e interespecies (porcentajes).

Taxonomía

Se realizó la clasificación filogenética de las proteínas codificadas en genomas completos mediante la creación de cladogramas utilizando el método UPGMA (*unweighted*

paired group means analysis).[20] A partir de los datos de cada secuencia *query*, se unieron las dos especies con mayor número de aciertos totales, luego se reformuló la tabla obteniendo promedios de los aciertos para las dos especies unidas, tomando el nuevo valor más alto y repitiéndolo hasta haber unido todas las especies *(figura 1)*. Finalmente se creó un cladograma consenso con base en los 8 creados anteriormente.

Técnica de análisis de datos:

El análisis de datos se realizó en tres procesos básicos: codificación, tabulación y construcción de tablas y gráficos. El analizador estadístico utilizado fue SPSS 20.0 en el cual se calculó el coeficiente de correlación de Spearman para establecer la correlación entre la proporción de *hits* de TaxPlot y el porcentaje de grupos funcionales de COGs por especie.

Limitaciones:

En principio, la presentación de especies analizadas no contiene la totalidad de enteropatógenos existentes, dado que COGs no incluía todos dentro de su base de datos. Segundo, se utilizó la cepa *Shigella dysenteriae* S197 en TaxPlot en sustitución de *Shigella dysenteriae* (cepa 1617) descrita en COGs. Tercero, no existen herramientas bioinformáticas para el diseño de cladogramas por el método UPGMA en valores obtenidos de TaxPlot. Por último, no se evaluaron patrones de función para cada COG de enterobacteria. Pese a estas limitaciones, nuestra investigación contribuye al desarrollo bioinformático aplicado a la microbiología clínica.

RESULTADOS

Se incluyó un total de 15 microorganismos (9 bacterias causantes de EDAs y 6 bacterias saprofitas [control]). El análisis de genomas para bacterias causantes de EDAs y control se muestra en los *cuadros II y III*, respectivamente.

11 de las 15 bacterias presentaron mayor COGs en las categorías J –12.25%– (6 bacterias causantes de EDAs y 5 controles). Para *Shigella dysenteriae*, *Escherichia coli* O157:H7 y *Klebsiella pneumoniae* el mayor número de COGs fue la categoría E –9.42%–, al igual que para la bacteria control saprofita *Escherichia coli* (cepa K-12, subcepa MG1655), no patógena *(cuadros II y III)*. Las siguientes categorías con mayor número de COGs en las especies analizadas son las categorías poco caracterizadas S y R (6.91 y 7.30% respectivamente), lo que podría indicar que algunas de las categorías que presentan pocos representantes en *clusters* podrían tener una mayor cantidad de representantes.

Las categorías funcionales para metabolismo fueron relativamente uniformes, la categoría I mantuvo proporciones proteicas para todos los microorganismos patógenos 3.26%, al contrario de la categoría G en la que las proporciones varían interespecies (6.45% total) *(cuadro III)*; en todos los demás casos se observaron fluctuaciones en las proporciones proteicas de *Enterococcus faecalis* (5.71, 7.53, 5.07, 5.14 y 4.72%, para las categorías C, E, F, H y P, respectivamente). Además, exceptuando éste, la proporción para las demás fue de 7.31% en la categoría C, 7.01% en P, 3.6% en F (excluyendo además a *Campylobacter jejuni* con 4.62%) y 6.61% en la categoría H (excluyendo además a *Morganella morganii* con 7.41%).

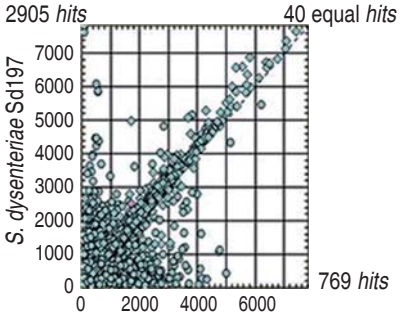
Conjuntamente, en la categoría funcional E se observó heterogeneidad de proporciones proteicas como *Campylobacter jejuni* y *Morganella morganii* (1.68%) con respecto a las demás bacterias patógenas (9.24%), excluyendo a *Enterococcus faecalis*. Por último, los valores del grupo metabólico Q fueron de 2.48% para *Klebsiella pneumoniae* y 1.36% para las demás bacterias causantes de EDAs *(cuadro III)*. Estos resultados ponen en relieve semejanzas con las proporciones de las bacterias control. En éstas, las proporciones de las categorías funcionales G (6.19%), P (5.38%), H (7.3%) y F (4.64%) varían heterogéneamente interespecies. *A contrario sensu* la categoría C (7.77% para *Escherichia coli*, 4.28% para *Lactobacillus plantarum* y 6.74% para las demás bacterias), la categoría E (6.82% para *Porphyromonas gingivalis*, 11.12% para *Veillonella parvula*, 10.14% para *Haemophilus influenzae* y 9.47% para las demás bacterias), la categoría I (5% para *Fusobacterium nucleatum* y 3.46% para las demás bacterias) y la categoría Q (proporciones bajas para todas las bacterias 1.20%) presentan variaciones para bacterias saprofitas específicas *(cuadro II)*.

Los grupos funcionales Y (0.00%), Z (0.03%), B (0.01%) y A (0.05%) fueron los de menor cuantía. Si bien ninguno presentó COGs dentro de la categoría Y (estructuras nucleares), algunas bacterias que mostraron COGs para la categoría funcional B (que involucra procesos de procesamiento y dinámica de cromatina) fueron *Veillonella parvula*, *Fusobacterium nucleatum*, *Klebsiella pneumoniae* y *Vibrio cholerae (cuadros II y III)*. En la *figura 2*, se muestra a manera de ejemplo la clasificación de los grupos funcionales para *Escherichia coli* O157:H7 (cepa Sakai) [Escc01], estos esquemas se realizaron en general para todas las bacterias incluidas en el estudio.

Las gráficas de líneas *(figura 3)* tuvieron como objetivo comparar las distribuciones de las categorías funcionales

Paso 1

Distribución of *E. coli* O157:H7 str. Sakal homologs



Enterica subsp. *Enterica* serovar *typhimurium*

Paso 2 Ciclo 1

		Sdy	Sen	Eco	Pmi	Efa	Kpn
Vibrio cholerae (Vch) (*Query proteins*: 3834)	Cje	39.18%	40.17%	40.30%	39.33%	30.91%	38.73%
	Sdy		63.38%	64.71%	59.39%	41.55%	61.71%
	Sen			66.38%	61.42%	43.04%	63.67%
	Eco				61.74%	42.91%	64.03%
	Pmi					41.55%	59.68%
	Efa						41.97%

Obtención de *hits* totales y creación de matriz de similitud

Unión de especies con mayor similitud

5316 query proteins produced 3714 *hits*

Ciclo 2

	Sdy	Eco + Sen	Pmi	Efa	Kpn
Cje	39.18%	40.23%	39.33%	30.91%	38.73%
Sdy		64.05%	59.39%	41.55%	61.71%
Eco + Sen			61.58%	42.97%	63.85%
Pmi				41.55%	59.68%
Efa					41.97%

Ajuste de la tabla promediando similitud de especies unidas con las demás. Unión con especie más similar

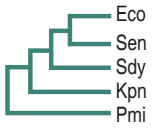
Ciclo 3

	Eco + Sen + Sdy	Pmi	Efa	Kpn
Cje	39.70%	39.33%	30.91%	38.73%
Eco + Sen + Sdy		60.49%	42.26%	62.78%
Pmi			41.55%	59.68%
Efa				41.97%

Se repite el proceso de ajuste por promedios y unión de especie más similar

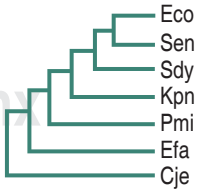
Ciclo 4

	Eco + Sen + Sdy + Kpn	Pmi	Efa
Cje	39.22%	39.33%	30.91%
Eco + Sen + Sdy + Kpn		60.08%	41.90%
Pmi			42.11%



Ciclo 5

	Eco + Sen + Sdy + Kpn + Pmi	Efa
Cje	39.28%	30.91%
Eco + Sen + Sdy + Kpn		42.01%



Se obtiene un cladograma basado en similitud para cada especie *query*.
Se obtiene árbol consenso a partir de éstos.

Figura 1. Flujograma para el desarrollo de cladograma con base en los datos del TaxPlot para las especies causantes de EDAs.

Cuadro II. Cantidad de COGs presentes en cada especie control analizada y porcentajes relativos al número total de COGs presentes en cada especie control.

	Haemophilus		*Escherichia*		*Fusobacterium*		*Lactobacillus*		*Veillonella*		*Porphyromonas*	
	Haeinf	%R	Esccol	%R	Fusnuc	%R	Lacpla	%R	Veipar	%R	Porgin	%R
J	178	13.78	202	9.35	166	13.82	167	12.78	166	13.38	163	15.45
A	1	0.08	2	0.09	0	0.00	0	0.00	0	0.00	0	0.00
K	54	4.18	104	4.81	55	4.58	78	5.97	51	4.11	44	4.17
L	84	6.50	107	4.95	74	6.16	81	6.20	77	6.20	70	6.64
B	0	0.00	0	0.00	1	0.08	0	0.00	2	0.16	0	0.00
D	27	2.09	38	1.76	20	1.67	29	2.22	25	2.01	19	1.80
Y	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
V	24	1.86	48	2.22	30	2.50	35	2.68	34	2.74	36	3.41
T	42	3.25	95	4.40	41	3.41	52	3.98	39	3.14	35	3.32
M	110	8.51	158	7.31	88	7.33	77	5.89	93	7.49	93	8.82
N	10	0.77	54	2.50	10	0.83	9	0.69	10	0.81	7	0.66
Z	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
W	8	0.62	9	0.42	7	0.58	3	0.23	6	0.48	1	0.09
U	30	2.32	42	1.94	24	2.00	16	1.22	27	2.18	20	1.90
O	82	6.35	106	4.91	49	4.08	49	3.75	54	4.35	57	5.40
X	32	2.48	16	0.74	1	0.08	25	1.91	3	0.24	6	0.57
C	85	6.58	168	7.77	82	6.83	56	4.28	86	6.93	70	6.64
G	83	6.42	166	7.68	71	5.91	106	8.11	58	4.67	46	4.36
E	131	10.14	203	9.39	110	9.16	129	9.87	138	11.12	72	6.82
F	54	4.18	80	3.70	50	4.16	70	5.36	61	4.92	58	5.50
H	85	6.58	137	6.34	93	7.74	71	5.43	109	8.78	100	9.48
I	48	3.72	72	3.33	60	5.00	49	3.75	35	2.82	39	3.70
P	72	5.57	161	7.45	57	4.75	72	5.51	53	4.27	50	4.74
Q	8	0.62	42	1.94	12	1.00	16	1.22	16	1.29	12	1.14
R	66	5.11	156	7.22	113	9.41	108	8.26	97	7.82	84	7.96
S	84	6.50	177	8.19	70	5.83	121	9.26	81	6.53	46	4.36
Total	1,398		2,343		1,284		1,419		1,321		1,128	

por grupos control y patogénico, interespecies y taxonómico, tanto en proporciones como en cantidades totales, permitiendo observar las diferencias en la distribución de genes en las categorías funcionales y las cantidades en que se presentan respectivamente.

Utilizamos TaxPlot como herramienta para observar la similitud entre las bacterias observadas, basado en su proteoma en lugar del alcance clásico que se enfoca en unas cuantas secuencias conservadas. Los *plots* generados por esta herramienta, algunos de ellos mostrados en la *figura 4*, nos brindan datos de aciertos o *hits* obtenidos al comparar todas las proteínas de un organismo *query* con las proteínas de otras dos especies seleccionadas mediante datos de BLAST precalculados, repartidos dependiendo del mayor parecido a una u otra especie en aciertos iguales o hacia una de éstas.

El análisis del TaxPlot sirvió además para la construcción de un cladograma con base en los datos de aciertos totales obtenidos en el mismo, de todas las combinaciones entre las 8 especies patogénicas con las que se crearon matrices de porcentajes y similitud *(cuadro IV)* y a partir de las cuales se creó un cladograma mediante el método UPGMA *(figura 5)*. Este cladograma, cuyo fundamento es la similitud del proteoma de las bacterias estudiadas, presenta gran similitud con las relaciones que existen para estos organismos basados en secuencias conservadas *(figura 5)*,[15,21] además, nos muestra la relación que existe entre nuestras especies respecto a sus proteomas,

Cuadro III. COG data base.
Cuadro III. Cantidad de COGs presentes en cada especie analizada y porcentajes relativos al número total de COGs presentes en cada especie.

	Campylobacter		*Shigella*		*Salmonella*		*Vibrio*		*Escherichia*		*Proteus*		*Enterococcus*		*Klebsiella*		*Morganella*	
	Camjej	%R	Shidys	%R	Salent	%R	Vibcho	%R	Escc01	%R	Promir	%R	Entfae	%R	Klepne	%R	Mormor	%R
J	158	14.03	194	9.67	204	9.24	194	9.90	202	8.98	196	10.17	167	11.77	198	8.78	197	10.42
A	0	0.00	2	0.10	2	0.09	1	0.05	2	0.09	1	0.05	0	0.00	1	0.04	1	0.05
K	37	3.29	99	4.94	104	4.71	78	3.98	105	4.67	89	4.62	84	5.92	104	4.61	88	4.66
L	62	5.51	105	5.23	103	4.66	99	5.05	109	4.84	100	5.19	95	6.69	105	4.65	94	4.97
B	0	0.00	0	0.00	0	0.00	1	0.05	0	0.00	0	0.00	0	0.00	1	0.04	0	0.00
D	19	1.69	38	1.89	38	1.72	38	1.94	38	1.69	37	1.92	23	1.62	38	1.68	37	1.96
Y	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
V	23	2.04	38	1.89	47	2.13	35	1.79	49	2.18	40	2.07	35	2.47	45	1.99	35	1.85
T	32	2.84	85	4.24	93	4.21	101	5.16	95	4.22	81	4.20	57	4.02	94	4.17	81	4.29
M	100	8.88	147	7.33	162	7.34	137	6.99	155	6.89	140	7.26	80	5.64	149	6.60	135	7.14
N	44	3.91	36	1.79	58	2.63	63	3.22	56	2.49	55	2.85	9	0.63	19	0.84	53	2.80
Z	1	0.09	0	0.00	0	0.00	1	0.05	0	0.00	1	0.05	0	0.00	1	0.04	1	0.05
W	2	0.18	10	0.50	12	0.54	16	0.82	10	0.44	11	0.57	3	0.21	12	0.53	8	0.42
U	28	2.49	40	1.99	58	2.63	52	2.65	62	2.76	77	3.99	19	1.34	58	2.57	41	2.17
O	65	5.77	102	5.08	108	4.89	105	5.36	110	4.89	97	5.03	51	3.59	110	4.88	95	5.03
X	2	0.18	22	1.10	43	1.95	17	0.87	67	2.98	44	2.28	30	2.11	26	1.15	35	1.85
C	86	7.64	149	7.43	168	7.61	137	6.99	168	7.47	147	7.62	81	5.71	165	7.31	151	7.99
G	34	3.02	137	6.83	173	7.84	118	6.02	159	7.07	108	5.60	122	8.60	176	7.80	100	5.29
E	124	11.01	196	9.77	200	9.06	175	8.93	204	9.07	177	9.18	104	7.33	213	9.44	194	10.26
F	52	4.62	71	3.54	80	3.62	69	3.52	79	3.51	70	3.63	72	5.07	77	3.41	75	3.97
H	78	6.93	131	6.53	149	6.75	127	6.48	138	6.13	127	6.59	73	5.14	155	6.87	140	7.41
I	39	3.46	63	3.14	68	3.08	64	3.27	69	3.07	61	3.16	50	3.52	73	3.24	64	3.39
P	77	6.84	159	7.93	145	6.57	126	6.43	159	7.07	135	7.00	67	4.72	171	7.58	127	6.72
Q	12	1.07	33	1.65	40	1.81	36	1.84	41	1.82	27	1.40	19	1.34	56	2.48	25	1.32
R	83	7.37	147	7.33	162	7.34	157	8.01	163	7.24	141	7.31	130	9.16	184	8.16	133	7.04
S	53	4.71	164	8.18	183	8.29	173	8.83	189	8.40	146	7.57	152	10.71	203	9.00	143	7.57
Total	1,211		2,168		2,400		2,120		2,429		2,108		1,523		2,434		2,053	

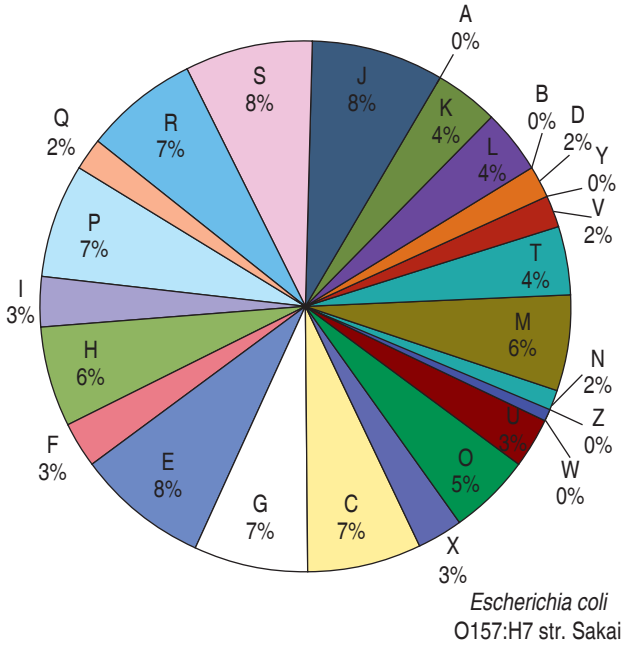


Figura 2. Lista de las categorías de la base de datos de COGs.

descubriendo como especies más cercanas a aquéllas que poseen un mayor número de proteínas similares.

Adicionalmente se realizó un cladograma basado en la secuencia de RNAr 16S de las especies analizadas, utilizando el programa Phylip versión 3.696 (*University of Washington*, http://evolution.genetics.washington.edu /phylip.html) y TreeView versión 1.6.6 (*University of Glasgow*, taxonomy.zoology.gla.ac.uk/rod/treeview.html) para poder comparar los métodos de alineamiento de secuencias con el que aquí se presenta basado en TaxPlot *(figura 5)*.

Las asociaciones entre COG no arrojaron resultados considerables. Se encontró una relación estadísticamente significativa entre la cantidad de proteínas de COG *data base* y TaxPlot (rho = 0.490; $p < 0.005$).

DISCUSIÓN

La coordinación interesante de los datos que se obtienen en estos análisis, en virtud de los cuales se logra apreciar las variables funcionales para cada microorganismo, que en un momento dado constituye un extraordinario interés

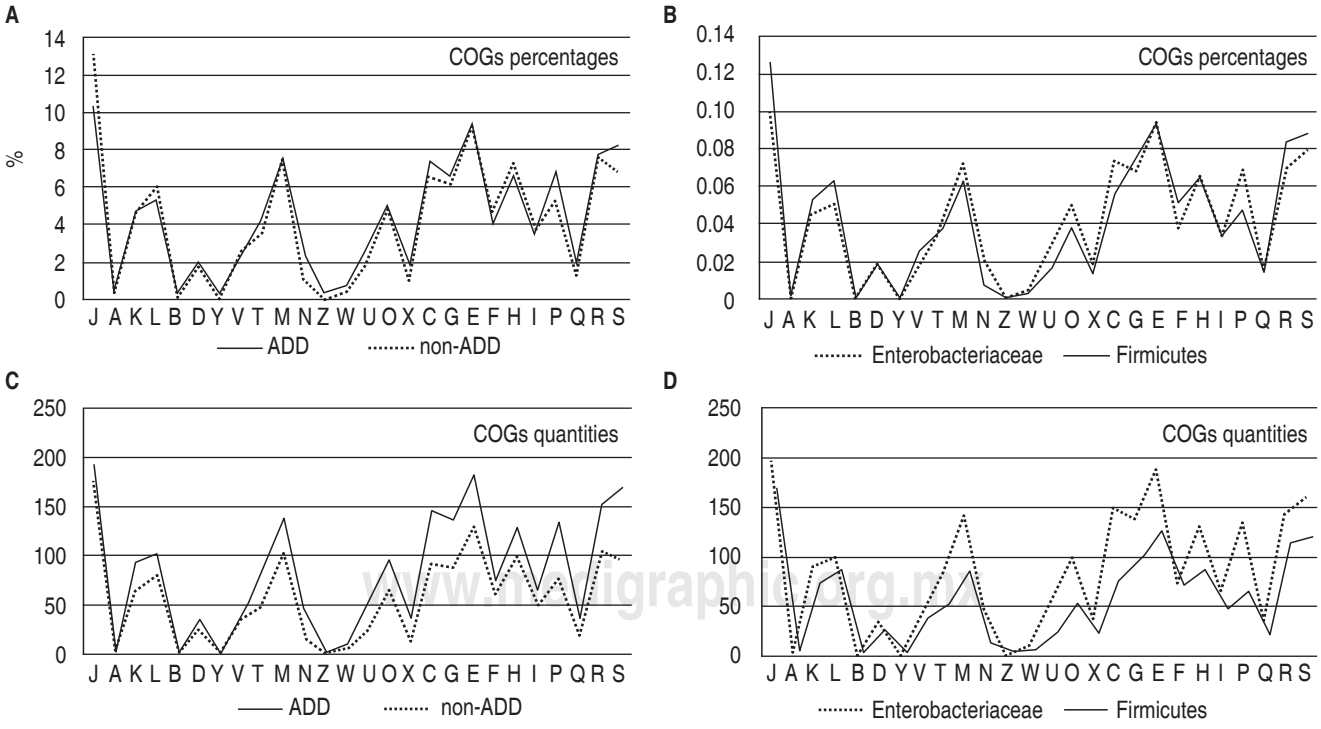


Figura 3. Comparaciones porcentual y de cantidades entre los grupos funcionales de COGs de bacterias patogénicas causantes de EDAs y saprofitas. **A y B.** Muestran proporciones considerando todas las bacterias; **C y D.** Muestran cantidades totales de COGs.

para el entendimiento de los mecanismos de patogenicidad, proporcionan resultados aproximados.

COG ***data base***. En el *cuadro III* se representó la distribución similar de COGs para cada especie en las categorías funcionales. Habiendo un mayor número de COGs en las categorías J y E, lo que es de esperarse para la categoría J que involucra procesos importantes en la síntesis de proteínas y biogénesis y para la categoría E vinculada en gran cuantía con la categoría J.[22] Aunque *Shigella dysenteriae*, *Escherichia coli* O157:H7 y *Klebsiella pneumoniae* presen-

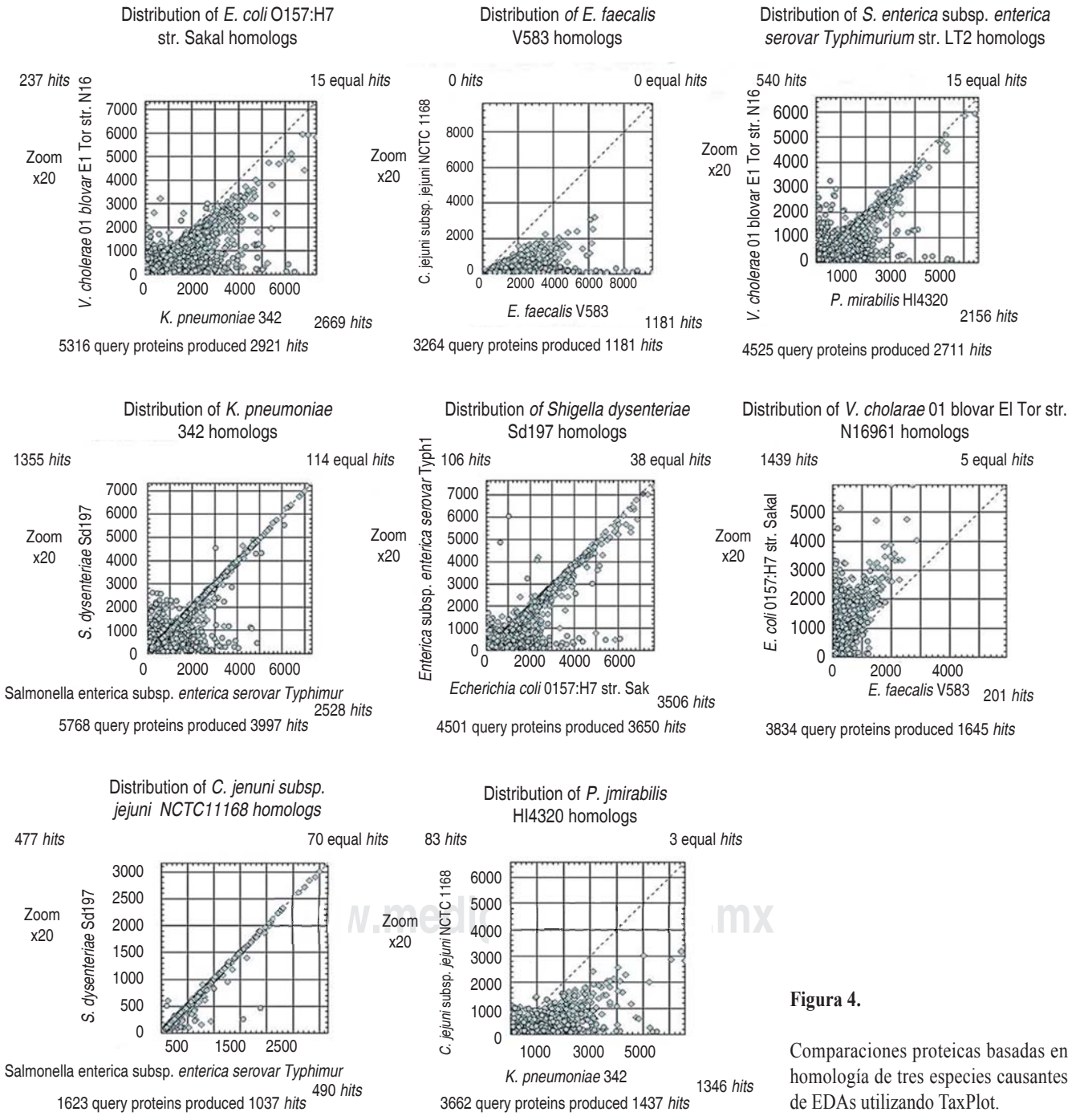


Figura 4.

Comparaciones proteicas basadas en homología de tres especies causantes de EDAs utilizando TaxPlot.

Cuadro IV. Porcentajes de las proteínas de las especies query que se obtuvieron como aciertos totales al compararlos con cada dos especies.

Especie query							
		Sen	Vch	Eco	Pmi	Efa	Kpn
Campylobacter jejuni (Cje)	Sdy	63.89%	59.46%	64.94%	61.61%	46.64%	61.37%
(Query proteins: 1623)	Sen		60.57%	65.87%	61.86%	47.26%	61.31%
	Vch			61.68%	59.21%	46.09%	57.42%
	Eco				63.40%	47.69%	62.35%
	Pmi					46.70%	59.58%
	Efa						46.52%
		Sen	Vch	Eco	Pmi	Efa	Kpn
Shigella dysenteriae (Sdy)	Cje	33.06%	31.17%	33.30%	32.02%	24.59%	32.28%
(Query proteins: 4501)	Sen		56.34%	81.09%	57.94%	37.55%	65.92%
	Vch			58.72%	51.74%	35.01%	55.05%
	Eco				60.32%	38.12%	68.25%
	Pmi					35.26%	55.77%
	Efa						35.73%
		Sdy	Vch	Eco	Pmi	Efa	Kpn
Salmonella enterica (Sen)	Cje	38.74%	37.08%	39.31%	38.17%	29.24%	37.72%
(Query proteins: 4525)	Sdy		62.52%	78.90%	68.20%	45.50%	73.68%
	Vch			64.80%	59.91%	43.43%	62.83%
	Eco				71.38%	47.47%	78.54%
	Pmi					43.91%	68.07%
	Efa						46.36%
		Sdy	Sen	Eco	Pmi	Efa	Kpn
Vibrio cholerae (Vch)	Cje	39.18%	40.17%	40.30%	39.33%	30.91%	38.73%
(Query proteins: 3834)	Sdy		63.38%	64.71%	59.39%	41.55%	61.71%
	Sen			66.38%	61.42%	43.04%	63.67%
	Eco				61.74%	42.91%	64.03%
	Pmi					41.55%	59.68%
	Efa						41.97%
		Sdy	Sen	Vch	Pmi	Efa	Kpn
Escherichia coli (Eco)	Cje	34.12%	34.42%	32.66%	33.56%	26.00%	33.43%
(Query proteins: 5316)	Sdy		69.86%	55.94%	61.49%	40.29%	66.42%
	Sen			56.64%	63.43%	41.22%	69.49%
	Vch				52.84%	37.62%	54.95%
	Pmi					38.58%	61.27%
	Efa						40.18%
		Sdy	Sen	Vch	Eco	Efa	Kpn
Proteus mirabilis (Pmi)	Cje	39.79%	40.44%	38.45%	40.66%	30.99%	39.10%
(Query proteins: 3662)	Sdy		69.47%	59.39%	70.62%	42.85%	66.11%
	Sen			61.58%	73.68%	44.40%	69.58%
	Vch				61.80%	41.81%	59.80%
	Eco					44.57%	70.32%
	Efa						43.39%

Continúa Cuadro IV. Porcentajes de las proteínas de las especies query que se obtuvieron como aciertos totales al compararlos con cada dos especies.

		Sdy	Sen	Vch	Eco	Pmi	Kpn
Enterococcus faecalis (Efa)	Cje	32.11%	32.51%	31.86%	33.21%	32.08%	32.35%
(Query proteins: 3264)	Sdy		49.23%	43.60%	50.86%	45.37%	47.92%
	Sen			45.31%	52.02%	47.33%	49.94%
	Vch				45.37%	43.20%	44.61%
	Eco					47.89%	50.37%
	Pmi						46.54%
		Sdy	Sen	Vch	Eco	Pmi	Efa
Klebsiella pneumoniae (Kpn)	Cje	35.30%	35.49%	34.28%	35.73%	35.00%	28.26%
(Query proteins: 5768)	Sdy		69.30%	58.88%	71.12%	61.91%	44.16%
	Sen			61.08%	73.79%	64.32%	45.89%
	Vch				61.13%	57.07%	42.63%
	Eco					64.96%	46.05%
	Pmi						43.24%

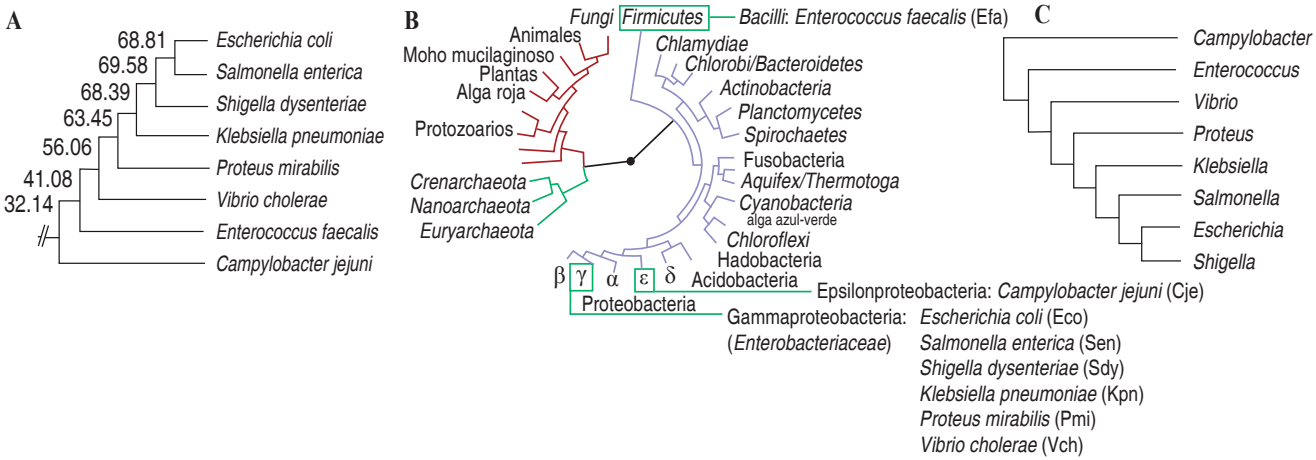


Figura 5. Cladogramas de comparativos. **A.** Árbol consenso obtenido con el método UPGMA sobre los valores de aciertos totales obtenidos con la herramienta bioinformática TaxPlot. **B.** Árbol filogenético donde se han señalado en verde los grupos estudiados pertenecientes a Proteobacteriae y Firmicutes mostrando divergencia evolutiva en torno a su ancestro en común (modificado de Ciccarelli FD, et al. [2006][40]), **C.** Cladograma basado en la secuencia de rRNA 16S de las especies analizadas utilizando el programa Phylip y TreeView.

tan elevaciones de la categoría E sobre la categoría J, las variaciones son mínimas sin poner de manifiesto alguna variable evolutiva que favorezca el desarrollo infectivo de alguna de ellas. Contrariamente, se observan diferencias en la categoría J para las bacterias utilizadas como control, en la que éstas presentan un mayor porcentaje para esta categoría, pero una menor cantidad de COGs que las bacterias causantes de EDAs. Para la categoría E observamos que las proporciones se mantienen similares aunque las bacterias causantes de EDAs presentan un mayor número de COGs, esto nos muestra la menor distribución de funciones hacia tareas generales en bacterias especializadas como las que se estudiaron, en las cuales se habría desarrollado un mayor número de genes dedicados a tareas relacionadas con su forma de vida *(cuadros II y III)*.

Las categorías S y R (8.14 y 7.66% respectivamente) corresponden a categorías poco caracterizadas, lo que indicaría que algunas de las categorías que presentan

pocos representantes podrían tener una mayor cantidad en otras categorías, principalmente en las relacionadas con los mecanismos de patogenicidad bacteriano causantes de EDAs, como T (transducción de señal, 4.15%) y V (mecanismos de defensa, 2.05%).

Al observar la categoría V en el *cuadro III* advertimos que la diferencia porcentual intergrupos bacterianos es leve, pero, en cuanto a las cantidades totales las bacterias causantes de EDAs tienen sólo unas pocas proteínas más que las demás bacterias. Algo similar puede advertirse en las categorías T y M *(cuadro II)*, aunque con notoria diferencia entre las bacterias patógenas que tienen casi el doble de proteínas. Para estas tres categorías observamos que las tres especies causantes de EDAs, que no son enterobacterias, poseen valores intermedios o cercanos a las bacterias control; tal variación sería lógica considerando que hemos utilizado como control bacterias que no son patógenas pero que siguen interaccionando con un hospedero (unidad holobionte) y aun cuando habrán aumentado en menor medida algunas defensas contra las del hospedero, aunque no en la cantidad que poseen las bacterias patógenas que han debido desarrollar estrategias como la resistencia múltiple a una gran cantidad de drogas, entre otras.[23,24] Estas diferencias pueden apreciarse de manera general en la *figura 3*, en la que se ha incluido también una comparación por grupos taxonómicos para poder diferenciar entre características que han aparecido en un grupo a través de la evolución, como se observa en la familia *Enterobacteriaceae* y otras que han surgido en varios grupos con similar estilo de vida, como las bacterias causantes de EDAs. En tales circunstancias, podríamos señalar que han desarrollado mecanismos semejantes relacionados con la patogenicidad, pudiendo haberse originado por convergencia evolutiva o por el fenómeno de transferencia genética horizontal de especies patógenas a especies no patógenas distantes que adquirieron patogenicidad al vivir en contacto con las que ya la poseían.[25-28]

Asimismo, las proteínas de metabolismo de estas bacterias fueron equivalentes. Los resultados demuestran que *Shigella dysenteriae*, *Escherichia coli* O157:H7 y *Salmonella enterica* subsp. *enterica* (bacterias heterótrofas y prototróficas) comparten entornos metabólicos similares para los COGs, coincidentes con los descritos por diversos autores.[29,30] Además, en este grupo bacteriano el catabolismo de carbono es crucial para su crecimiento y proliferación dentro de células huéspedes de mamíferos, relacionado con más de un COGs *(figura 3)*.[31] Exceptuando la categoría E, distinguimos que las categorías C, G, H y P responsables del metabolismo energético, de carbohidratos, de iones y de coenzimas respectivamente, presentan las mayores proporciones de COGs en todos los organismos y en mayores cantidades, aunque en proporciones similares, para los casos de bacterias causantes de EDAs, con diferencias muy marcadas para las bacterias pertenecientes a *Enterobacteriaceae (figura 3 y cuadro III)*. Las dos primeras categorías nos indican el rol e importancia de obtener energía y las otras dos, de la necesidad de regular la captación y uso de iones y coenzimas en el entorno en que se desarrollan estas bacterias, exigencias requeridas para su bienestar y que están en constante proceso, aunque de ritmo distinto entre microorganismos.[32]

La categoría F (transporte y metabolismo de nucleótidos) que se mantiene casi constante en todos los grupos *(figura 3)*, representa el metabolismo básico de las bacterias. Podría esperarse un aumento similar al observado en las categorías anteriores, ya que son bacterias que se reproducen rápidamente respecto al grupo de bacterias saprófitas utilizadas como control. Una explicación lógica podría ser la adaptación que tienen las bacterias causantes de EDAs a diferentes microambientes, lo que supondría un mayor número de genes orientados a diferentes tareas metabólicas dependientes del microambiente en que se encuentran,[33,34] lo cual explicaría también el porqué estas bacterias tienen una cantidad de genes tan grande respecto a otros grupos.

En contraposición, al realizar la comparación con la cantidad total de COGs, se observan diferencias muy marcadas principalmente porque las bacterias causantes de EDAs poseen de 1.5 hasta casi dos veces más COGs que las bacterias saprofitas *(figura 3, cuadros II y III)*. Las bacterias control en las que no se encuentran diferencias (5.60% ± 0.09 de proteínas), en tales casos aunque nimios en apariencia, la cantidad de COGs manifiesta probablemente la misma cantidad de proteínas funcionales para el metabolismo. La diferencia radicaría en cómo influyen estas proteínas en el *fitness*, la infectividad, virulencia, etc.; podemos ejemplificar este fenómeno como una caja COGs de herramientas para cada bacteria, de las que tendrá que hacer uso dependiendo del hospedero, las condiciones, el rendimiento, etc., en ese sentido cuando se evaluó el proteoma divergente entre *Escherichia coli* O157:H7 y *Escherichia coli* (cepa K-12), se encontraron rasgos distintivos mínimos (103 y 17 proteínas diferenciales por COG, respectivamente; la mayor proporción elevada fue la categoría X (2.98 y 0.74%, respectivamente), relacionados con el mobiloma con extraordinario interés científico que apoyados en el mismo principio pueden resaltar datos sobre virulencia, propagación, transmisión genética horizontal o rango de hospedero restringido a humanos),[35,36] resultado que diverge de los análisis realizados en otras enterobacterias.[37,38]

Es de notarse que una gama extensa de pesquisas pueden llevarse a cabo en todos los COGs para cada bacteria, no es objetivo del estudio ni es esta la ocasión para enumerar ni

describir cada una de ellas, sin embargo, una consideración no puede pasar inadvertida: 2 de las 9 bacterias patogénicas (*Klebsiella pneumoniae* y *Vibrio cholerae*) y 2 de las 6 bacterias control (*Fusobacterium nucleatum* y *Veillonella parvula*) presentaron COGs para la categoría B (estructura y dinámica de cromatina), lo cual es extraño de observar en bacterias y requiere atención especial. Este COG, presente en ambas bacterias patógenas corresponde a una desacetilasa de histona, la cual consideramos puede estar relacionada con la regulación en la expresión de la virulencia de estas bacterias en presencia de condiciones desfavorables que se da a nivel del nucleoide de estas bacterias, el cual cumple una función similar al de las histonas en *Eukarya*.[39]

TaxPlot: con ayuda de la herramienta bioinformática TaxPlot del NCBI se logró acceder a datos de similitud entre todo el proteoma de las bacterias analizadas en grupos de a tres, en la *figura 4* se aprecian a manera de ejemplo, algunos *plots* obtenidos con esta herramienta de un total de 168, en los que puede observarse la similitud entre cada especie seleccionada como *query* con otras dos especies de todo el grupo de bacterias. A partir de estos datos porcentualizados obtenidos de los aciertos totales, se creó un cladograma *(figura 5)* cuya metodología de elaboración se describió en el flujograma de la *figura 1*. Este cladograma basado en similitud del proteoma bacteriano presenta gran similitud con las relaciones que existen para estos organismos basados en secuencias conservadas *(figura 5)*.[15,19,21,40] En la comparación entre cladogramas (TaxPlot, alineamiento de familias universales de proteínas y RNAr 16S) las bacterias analizadas se agrupan en general de manera muy similar. *Campylobacter jejuni* mostró un carácter plesiomórfico. Nótese la separación entre *Enterobacteriaceae* (*Salmonella enterica*, *Escherichia coli* O157:H7, *Shigella dysenteriae*, *Proteus mirabilis* y *Klebsiella pneumoniae*) y *Vibrionaceae* (*Vibrio cholerae*); además, entre éstos (gammaproteobacteria*)* con epsilonproteobacteria *(Campylobacter jejuni)* y *Firmicutes* (*Enterococcus faecalis*), en los que la similitud entre enterobacterias oscila entre 63.45 y 68.81% con el método basado en TaxPlot, sin embargo, al ir incorporando bacterias de otros *clados* esta similitud decrece drásticamente, a 56.06 para *Vibrio cholerae*, 41.08 para *Enterococcus faecalis* y 32.14 para *Campylobacter jejuni*.

La única excepción a la similitud de resultados entre el método basado en TaxPlot y el basado en secuencias rRNA 16S es la relación entre *Shigella dysenteriae*, *Salmonella enterica* y *Escherichia coli* O157:H7, en la que obtenemos a *Escherichia coli* y *Salmonella enterica* más relacionadas entre ellas que con *Shigella dysenteriae*, cuando lo usual es observar a *Escherichia coli* más relacionada con *Shigella dysenteriae* que con *Salmonella enterica (figura 5)*.

Este método parece tener como limitantes la diferencia en el número de proteínas de la especie *query*, ya que *Campylobacter jejuni* presenta el menor número (1,623), en comparación con números mayores a 3,000 en las otras bacterias, de tal suerte que el bajo número de coincidencias la «muestra» como la más alejada, mientras que por métodos basados en familias de proteínas conservadas[40] se ha demostrado que esta bacteria debería estar unida a las demás proteobacterias y *Enterococcus faecalis* tendría que aparecer más alejada, ya que pertenece al filo *Firmicutes*.[22] Curiosamente, al alinear y obtener un cladograma a partir del ARN ribosomal 16S, conservamos el mismo orden entre estas dos bacterias respecto al conseguido comparando el proteoma *(figura 5)*.

El método que presentamos en este estudio representa una aproximación holística al estudio de las relaciones filogenéticas entre especies cercanas que podría contrastarse con otras aproximaciones más específicas para poder observar el comportamiento evolutivo de organismos que comparten un mismo ambiente o nicho. En el caso de bacterias, se sabe que han pasado por varios eventos de transferencia genética horizontal haciendo más compleja la identificación de la historia evolutiva en estos grupos, ya que un gen ortólogo puede ser más similar en un organismo que no necesariamente sea el más cercano evolutivamente,[41,42] al realizar una comparación entre todo el proteoma puede disminuirse un poco este efecto. El presente método reduce el tiempo de análisis filogenético y establece una reconstrucción filogenética a nivel proteico; tales características funcionan siempre y cuando no se escojan microorganismos muy distantes que actúen como *outliers* y disminuyan la aplicación del método.

Estos elementos de juicio pueden transportarse en datos exactos empleando procedimientos bioinformáticos que puedan como en este caso, describir las características de los agentes patógenos productores de la enfermedad.

AGRADECIMIENTOS

A Manuel A. Mendoza Aquije MSc, Carla Gallo López-Aliaga, MSc y Mirko Zimic Peralta, PhD, mentores, colaboradores y amigos que nos contagiaron su vocación por la bioinformática y el desarrollo científico.

CONFLICTO DE INTERESES

Los autores declaran no tener conflicto de intereses.

FINANCIAMIENTO

Autofinanciado por los autores.

REFERENCIAS

1. MINSA. Plan de comunicaciones. Prevención de enfermedades diarreicas agudas (EDA) y cólera 2014. Versión preliminar. Disponible en: http://www.minsa.gob.pe/portada/Especiales/2014/lavadomanos/archivo/Plan_de_comunic aciones-prevencion_de_enfermedades_diarreicas_y_colera.pdf
2. World Health Organization. The world health report: making a difference. Geneva: World Health Organization; 1999.
3. Moya SJ, Pio DL, Terán VA, Olivo LJ. Yield diagnosis of blood agar with filter against karmali agar for isolation of *Campylobacter* in stool culture. J Microbiol Meth. 2015; [en prensa].
4. Winn W, Allen S, Janda W, Koneman E, Procop G, Schreckenberger P et al. Koneman's color atlas and textbook of diagnostic microbiology. 6th ed. Philadelphia: Lippincott Williams & Wilkins; 2006.
5. Murga H, Huicho L, Guevara G. Acute diarrhoea and *Campylobacter* in Peruvian children: a clinical and epidemiologic approach. J Trop Pediatr. 1993; 39 (6): 338-341.
6. Seas C, Alarcón M, Aragón JC, Beneit S, Quiñonez M, Guerra H et al. Surveillance of bacterial pathogens associated with acute diarrhea in Lima. Int Infect Dis. 2000; 4 (2): 96-99.
7. Cama RI, Parashar UD, Taylor DN, Hickey T, Figueroa D, Ortega YR et al. Enteropathogens and other factors associated with severe disease in children with acute watery diarrhea in Lima, Peru. J Infect Dis. 1999; 179 (5): 1139-144.
8. Roman LT, Michael YG, Darren AN, Eugene VK. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 2000; 28 (1): 33-36.
9. Thorat SS, Thakare VP. Analysis of *Staphylococcus* using comparative genomics. IJSER. 2013; 4 (12): 1775-1779.
10. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2005; 33 (Database issue): D39-D45.
11. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science. 1997; 278 (5338): 631-637.
12. Roman LT, Michael YG, Darren AN, Eugene VK. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 2000; 28 (1): 33-36.
13. Wheeler D, Benson D, Rapp B. Using TaxPlot to compare genomes. Bethesda: NCBI News; 2001: p. 1-2.
14. Makarova SK, Wolf IY, Koonin VE. Archaeal clusters of orthologous genes (arCOGs): an update and application for analysis of shared features between thermococcales, methanococcales and methanobacteriales. Life. 2015; 5: 818-840.
15. Neupane S, Finlay RD, Alström S, Goodwin L, Kyrpides NC, Lucas S et al. Complete genome sequence of *Serratia plymuthica* strain AS12. Stand Genomic Sci. 2012; 6: 165-173.
16. Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML, Dodson RJ et al. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. Nature. 2000; 406 (6795): 477-483.
17. Bhagwat AA, Bhagwat M. Methods and tools for comparative genomics of food borne pathogens. Foodborne Pathog Dis. 2008; 5 (4): 487-497.
18. Siddaramappa SS. Comparative and functional genomic studies of *Histophilus somni* (*Haemophilus somnus*) [Thesis]. Blacksburg: Virginia Polytechnic Institute and State University; 2007.
19. Montiel AF. Flora bacteriana habitual. Boletín Escuela de Medicina U.C. Pontificia Universidad Católica de Chile. 1997; 26: 133-139.
20. Kuske RC, Ticknor LO, Miller ME, Dunbar JM, Davis JA, Barns SM et al. Comparison of soil bacterial communities in rhizospheres of three plant species and the interspaces in an arid grassland. Appl Environ Microbiol. 2002; 68 (4): 1854-1863.
21. Moran AN, Russell JA, Koga R, Fukatsu T. Evolutionary relationships of three new species of *Enterobacteriaceae* living as symbionts of aphids and other insects. Appl Environ Microbiol. 2005; 71 (6): 3302-3310.
22. Chieh-Hua L, Chun-Yi L, Chao AH, Feng-Chi C. Changes in transcriptional orientation are associated with increases in evolutionary rates of enterobacterial genes. BMC Bioinformatics. 2011; 12 (Suppl 9): S19.
23. Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J et al. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. Nature. 2001; 413 (6858): 848-852.
24. Salvucci E. Selfishness, warfare, and economics; or integration, cooperation, and biology. Cell Infect Microbiol. 2012; 54 (2): 1-12.
25. Huddleston RJ. Horizontal gene transfer in the human gastrointestinal tract: potential spread of antibiotic resistance genes. Infect Drug Resist. 2014; 7: 167-176.
26. Petty NK, Bulgin R, Crepin VF, Cerdeño-Tárraga AM, Schroeder GN, Quail MA et al. The *Citrobacter rodentium* genome sequence reveals convergent evolution with human pathogenic *Escherichia coli*. J Bacteriol. 2010; 192 (2): 525-538.
27. Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M et al. The complete genome sequence of *Escherichia coli* K-12. Science. 1997; 277 (5331): 1453-1462.
28. Perna NT, Plunkett G 3rd, Burland V, Mau B, Glasner JD, Rose DJ et al. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature. 2001; 409 (6819): 529-533.
29. Barabote RD, Saier MH Jr. Comparative genomic analyses of the bacterial phosphotransferase system. Microbiol Mol Biol Rev. 2005; 69: 608-634.
30. Bell KS, Sebaihia M, Pritchard L, Holden MTG, Hyman LJ, Holeva MC et al. Genome sequence of the enterobacterial phytopathogen *Erwinia carotovora* subsp. *atroseptica* and characterization of virulence factors. Proc Natl Acad Sci USA. 2004; 101: 11105-11110.
31. Götz A, Eylert E, Eisenreich W, Goebe W. Carbon metabolism of enterobacterial human pathogens growing in epithelial colorectal adenocarcinoma (Caco-2) cells. PLoS One. 2010; 5 (5): 1-14.
32. Emerson AE. Dynamic homeostasis: a unifying principle in organic, social and ethical evolution. Zygon. 1968; 3 (2): 129-168.
33. Alteri CJ, Mobley HLT. *Escherichia coli* physiology and metabolism dictates adaptation to diverse host microenvironments. Curr Opin Microbiol. 2012; 15 (1): 3-9.
34. Luchi S, Weiner L. Cellular and molecular physiology of *Escherichia coli* in the adaptation to aerobic environments. J Biochem. 1996; 120: 1055-1063.
35. Barkay T, Smets BF. Horizontal gene flow in microbial communities. ASM News. 2005; 71: 412-419.
36. Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol. 2005; 3: 722-732.
37. Sabbagh SC, Forest CG, Lepage C, Leclerc JM, Daigle F. So similar, yet so different: uncovering distinctive features in the genomes of *Salmonella enterica* serovars *typhimurium* and Typhi. FEMS Microbiol Lett. 2010; 305 (1): 1-13.
38. Welch AR, Burland V, Plunkett G, Redford P, Roesch P, Rasko D et al. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci USA. 2002; 99 (26): 17020-17024.
39. Ghosh A, Paul K, Chowdhury R. Role of the histone-like nucleoid structuring protein in colonization, motility and bile dependant repression of virulence gene expression in *Vibrio cholerae*. Infect Immun. 2006; 74 (5): 3060-3064.
40. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. Science. 2006; 311 (5765): 1283-1287.
41. Stecher B, Denzler R, Maier L, Bernet F, Sanders MJ, Pickard DJ, et al. Gut inflammation can boost horizontal gene transfer between pathogenic and comensal *Enterobacteriaceae*. PNAS. 2012; 109 (4): 1269-1274.
42. Akortha EE, Filgona J. Transfer of gentamicin resistance genes among *Enterobacteriaceae* isolated from the out patents with urinary tract infections attending 3 hospitals in Mubi, Adamawa State. Sci Res Essays. 2009; 4 (8): 745-752.