

1996B - 2002B

093671996

UNIVERSIDAD DE GUADALAJARA

CENTRO UNIVERSITARIO DE CIENCIAS
BIOLÓGICAS Y AGROPECUARIAS
DIVISIÓN DE CIENCIAS BIOLÓGICAS Y AMBIENTALES



**Identificación y análisis del proceso evolutivo
de los genes transferidos horizontalmente al
ancestro común de *Escherichia* y *Salmonella***

**TRABAJO DE TITULACIÓN EN LA MODALIDAD DE
TESIS**

**QUE PARA OBTENER EL TÍTULO DE
LICENCIADO EN BIOLOGÍA**

**PRESENTA
SANTIAGO CASTILLO RAMÍREZ**

LAS AGUJAS ZAPOPAN, JALISCO. DICIEMBRE DE 2002



UNIVERSIDAD DE GUADALAJARA

CENTRO UNIVERSITARIO DE CIENCIAS BIOLÓGICAS Y AGROPECUARIAS

COORDINACIÓN DE CARRERA DE LA LICENCIATURA EN BIOLOGÍA

COMITÉ DE TITULACIÓN

**C. SANTIAGO CASTILLO RAMÍREZ
PRESENTE.**

Manifestamos a Usted que con esta fecha ha sido aprobado su tema de titulación en la modalidad de **TESIS E INFORMES** opción Tesis con el título "IDENTIFICACIÓN Y ANÁLISIS DEL PROCESO EVOLUTIVO DE LOS GENES TRANSFERIDOS HORIZONTALMENTE AL ANCESTRO COMÚN DE *Escherichia* y *Salmonella*", para obtener la Licenciatura en Biología.

Al mismo tiempo le informamos que ha sido aceptado/a como Director de dicho trabajo el/la **DR. ALEJANDRO GARCIRRUBIO GRANADOS** y como Asesor el/la **DRA. ANNE SANTERRE LUCAS**.

**A T E N T A M E N T E
"PIENSA Y TRABAJA"**

**"2002, Año Constanancio Hernández Alvirde"
Las Agujas, Zapopan, Jal., 21 de noviembre del 2002**



**DRA. MÓNICA ELIZABETH ROJAS LÓPEZ
PRESIDENTE DEL COMITÉ DE TITULACIÓN**

COORDINACIÓN DE LA CARRERA DE
LICENCIADO EN BIOLOGÍA

Leticia Hernández López

**M.C. LETICIA HERNÁNDEZ LÓPEZ
SECRETARIO DEL COMITÉ DE TITULACIÓN**

c.c.p. **DR. ALEJANDRO GARCIRRUBIO GRANADOS**.-Director del Trabajo.

c.c.p. **DRA. ANNE SANTERRE LUCAS**.-Asesor del Trabajo.

c.c.p. Expediente del alumno

MERL/LHL/mam

199252
B750
97

C. DRA. MÓNICA ELIZABETH RIOJAS LOPEZ
PRESIDENTE DEL COMITÉ DE TITULACIÓN
DE LA DIVISIÓN DE CIENCIAS BIOLÓGICAS Y AMBIENTALES
DE LA UNIVERSIDAD DE GUADALAJARA
PRESENTE

Por medio de la presente, nos permitimos informar a usted, que una vez revisado el trabajo de TESIS que realizó el pasante SANTIAGO CASTILLO RAMÍREZ con el título IDENTIFICACIÓN Y ANÁLISIS DEL PROCESO EVOLUTIVO DE LOS GENES TRANSFERIDOS HORIZONTALMENTE AL ANCESTRO COMÚN DE *ESCHERICHIA* Y *SALMONELLA* consideramos que ha quedado debidamente concluido, por lo que ponemos a su consideración el escrito final para autorización de impresión y en su caso programación de fecha de exámenes de tesis y profesional respectivos.

Sin otro particular, agradecemos de antemano la atención que sirva brindar a la presente y aprovechamos la ocasión para enviarle un cordial saludo.

ATENTAMENTE

Las Agujas, Zapopan, Jal., a 28 de Noviembre del 2002

DIRECTOR DEL TRABAJO

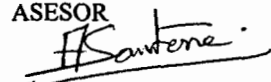

NOMBRE Y FIRMA

DR. ALEJANDRO GARCÍARRUBIO GRANADOS



COORDINACIÓN DE LA CARRERA DE
LICENCIADO EN BIOLOGÍA

ASESOR



NOMBRE Y FIRMA

ANNE SANTERRE LUCAS

1./DR. CARLOS ALVAREZ MOYA
NOMBRE COMPLETO

FIRMA



Patricia Castro

2./M.C. PATRICIA CASTRO FELIX
NOMBRE COMPLETO

FIRMA

3./DR. AARÓN RODRÍGUEZ CONTRERAS
NOMBRE COMPLETO

FIRMA


FIRMA

S./DR. DANIEL ORTUÑO SAHAGUN
NOMBRE COMPLETO

FIRMA

**Tesis realizada en el Departamento de Reconocimiento Molecular y Bioestructura,
INSTITUTO DE BIOTECNOLOGIA,
UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
Cuernavaca, Morelos.**

Director: Dr. Alejandro Garciarribio Granados.

Asesora: Dra. Anne Santerre Lucas.

AGRADECIMIENTOS

A mi Madre por su inmarcesible alegría, por su inagotable esperanza; por ese saco de tiempo que hay que llenar.

A Urbanito por ser el paladín de la intensidad y el arrojo, por lo rotundo de sus palabras e ideas; por compartir los cronopios y las famas.

A Patotas porque sin ti manita no sería manita.

A la familia burron por la memoria compartida.

A la Niña Morena porque un día me dijo que me quería. Gracias a ti se que en el esfuerzo se funden el corazón y la razón.

A mis tías por su dedicación y cariño.

A Alejandro por enseñarme a ser autodidacta pero sobre todo por la confianza.

A los compas, Macaco y Capitán Pitorcas, por las cascaras, las chelas; por su amistad.

A los profesores por aguantarme.

A la U. de G. por las oportunidades.

A la memoria de Cristóbal Castillo, Gonzalo Herrera, Fernando Alfaro, Olaya Ramírez y Jaime Miranda.

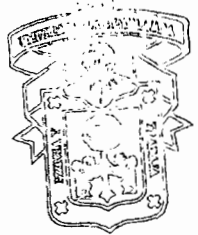
En otoño caen las hojas, en invierno las pisamos.

RESUMEN

La transferencia horizontal de genes es un fenómeno de gran importancia en la evolución de las bacterias. Con el fin de determinar cuál es el proceso evolutivo de los genes transferidos horizontalmente, se desarrolló y aplicó un modelo que permitió la localización de los genes transferidos horizontalmente al ancestro común de los géneros *Escherichia* y *Salmonella*; y se analizó el tipo de selección que actuaba sobre dichos genes. Según el modelo empleado, fueron nueve los genes localizados como transferidos. Para cuatro de ellos genes se pudo establecer los cladogramas de los posibles donadores. Las distancias entre dichos cladogramas y el cladograma de *Escherichia* y *Salmonella* tuvieron un amplio rango: desde un clado en la misma división, la de las Proteobacterias -que implica cientos de millones de años-, hasta un clado en el dominio de las arqueas -que implica miles de millones de años-. El análisis de los nueve genes sugiere que el tipo de selección predominante es la presión selectiva purificadora. Pero si nos atenemos a la variabilidad de la selección en los sitios de las proteínas de los genes en cuestión, hubo dos grupos bien diferenciados: uno en el que la presión selectiva negativa es intensa y es el único elemento, comprendidos todos sus datos en el rango que va de 0 a 0.6 -para los valores de w -; otro en el que la presión selectiva negativa es principal componente, aunque es laxa y de menor intensidad. Para este último grupo hay una parte considerable de los datos -más del 5 %- que es explicada por procesos similares al neutralismo y por presión selectiva positiva. Al parecer no hay diferencias entre el grupo de los genes transferidos y el grupo control, ya que en este último también la presión selectiva negativa es el principal componente. Cuando se toma en consideración la variabilidad de la selección a lo largo de los sitios de la proteína, el grupo control tiene subgrupos similares a los descritos para los genes transferidos horizontalmente.

INDICE

INTRODUCCIÓN.....	1
ANTECEDENTES.....	3
PLANTEAMIENTO.....	4
OBJETIVOS.....	5
ESTRATEGIA EXPERIMENTAL.....	6
MATERIAL Y MÉTODOS	
Material.....	7
Metodología.....	11
RESULTADOS	
Localización de genes.....	13
Valor promedio de la selección.....	15
Variabilidad de la selección en los linajes.....	16
Variabilidad de la selección en los sitios.....	17
DISCUSIÓN.....	21
CONCLUSIONES.....	25
ANEXO 1.....	26
ANEXO 2.....	28
APÉNDICE.....	30
BIBLIOGRAFÍA.....	33
GLOSARIO.....	35



BIBLIOTECA CENTRAL

INTRODUCCIÓN

Diferencias en el tamaño del genoma entre las especies bacterianas reflejan una variación del acervo genético. Ellas se deben principalmente a la adquisición y pérdida de genes (1). El análisis de secuencias ha proporcionado evidencias de transferencia lateral de genes, decaimiento genético y variación de la expresión genética en diferentes patógenos microbianos (2)

Uno de los mecanismos más importantes en la adquisición de material genético es la transferencia horizontal de genes. Otro es la generación interna de material genético, por medio del fenómeno de duplicación (paralogía). La transferencia horizontal es un evento frecuente en las bacterias y ocurre incluso entre organismos que tienen grandes distancias filogenéticas de por medio (3,4). A través de ella se pueden introducir capacidades metabólicas completamente funcionales que hacen posible la efectiva explotación de nuevos ambientes y, con ello, la diversificación y la especiación bacteriana (5). Así, gracias al gen adquirido, la transferencia horizontal lleva a los organismos a ser buenos competidores en nichos ecológicos previamente inexplotables (3). Si bien antes se creía que la transferencia era un hecho irrefutable, se pensaba asimismo que era un evento poco frecuente e irrelevante frente a la transferencia vertical de los genes. Actualmente se sabe que en algunas especies la transferencia horizontal tiene tanta importancia como la transferencia vertical. La transferencia horizontal de genes y su subsecuente incorporación en el genoma receptor es una fuerza central que dirige la evolución de las bacterias (17).

Aunque el árbol de la vida es útil para recrear el contexto evolutivo, también es objeto de muchas controversias. Quizás la mayor de las cuales radica en si un árbol basado en un solo gen -subunidad pequeña del RNAr- puede representar la evolución de las especies. El uso de un solo árbol asume que las especies están relacionadas a través de la descendencia vertical. Sin embargo, no todos los genes siguen las reglas de la descendencia vertical. Algunos pueden ser transferidos entre linajes por transferencia horizontal. Este fenómeno complica la reconstrucción evolutiva, porque implica que

algunas especies son quiméricas, con diferentes historias, para diferentes partes de su genoma (11).

La descendencia horizontal de genes entre diferentes especies es un fenómeno evolutivo en discusión, sobre todo cuando se considera la transferencia entre imperios - bacterias, arqueas y eucariotes-. Este fenómeno parece cuestionar dos puntos de vistas imperantes en biología: 1) el que la evolución de la vida pueda ser representada por un árbol -descendencia vertical- donde todo el material genético de un organismo proviene de sus antecesores; y, 2) el del aislamiento reproductivo entre las especies. Respecto al primer punto, si se toman en cuenta el proceso de transferencia horizontal, las ramas del árbol de la vida no serían linajes aislados, sino que habría continuos puentes entre las mismas. En lo que toca al segundo punto, el material genético que definiría a una especie no estaría del todo aislado, habría material procedente de descendencia vertical y material adquirido por transferencia horizontal, por ende no habría aislamiento reproductivo (pues organismos de incluso diferentes imperios que compartirían genes que fueron transferidos).

La transferencia horizontal tiene un papel preponderante en la diversificación y especiación de las especies bacterianas. Las primeras aproximaciones basadas en los análisis de mejores puntuaciones para taxones específicos indican un alto nivel de transferencia horizontal para la mayoría de los genomas bacterianos y de arqueas. Más que ser un fenómeno raro o atípico, parece que es un proceso común (7).

Desde una perspectiva teórica-evolutiva, la transferencia horizontal, particularmente cuando ocurre entre eucariotes y bacterias, es decir, entre imperios, es un testimonio de la notable unidad de los mecanismos biológico-moleculares, que resulta en la compatibilidad de proteínas eucarióticas y bacterianas que han evolucionado en diferentes medios a lo largo de millones de años (10). Aunque suele ser más abundante la transferencia horizontal entre los genes operacionales (7) -los implicados en el metabolismo-, ella se produce frecuentemente en los genes informacionales -los implicados en el procesamiento de la información- (6,10), por lo cual al parecer no hay genes intocables que escapen al fenómeno de la transferencia horizontal.

ANTECEDENTES

La transferencia horizontal de genes entre diferentes especies está bien documentada en las enterobacterias patógenas, en las que es un mecanismo importante en la diversificación de la virulencia y en el traspaso de genes de resistencia a los antibióticos. El ácido desoxirribonucleico (DNA) puede ser transferido por varios mecanismos: transformación, conjugación y transducción (6). La transferencia generalmente involucra cassettes de genes que oscilan entre 5 y 100 kilobases.

El imperio de las bacterias es el que cuenta con mayor número de especies secuenciadas. La división de las Proteobacterias es la mejor representada, ya que tiene representantes de todas las subdivisiones -alfa, beta, gama, delta, épsilon-. La subdivisión gama cuenta con muchos organismos secuenciados; de entre ellos la familia Enterobacteriaceae cuenta con tres géneros: *Escherichia*, *Salmonella* y *Yersinia*. Incluye además miembros de las siguientes familias; Pasteurellaceae, Pseudomonadaceae, Vibrionaceae. Por tanto, debido a que el clado de la subdivisión gamma está bien representado por las especies actualmente secuenciadas, se escogió el ancestro común de *Escherichia* y *Salmonella* para localizar las transferencias horizontales.

En 1998 Lawrence y Ochman encontraron que de los 4288 marcos de lectura abierta 755 han sido introducidos en el genoma de *Escherichia coli*; en al menos 234 eventos de transferencia horizontal desde que esta especie divergió del linaje de *Salmonella*, 100 millones de años atrás. El promedio del tiempo de traspaso de los genes fue 14.4 millones de años, lo cual da una tasa de transferencia de 16Kb por millón de años en este linaje desde la divergencia. Sin embargo, aunque la mayoría de los genes adquiridos fueron perdidos, las secuencias que persistieron -18% del actual cromosoma de *E. coli*-, proporcionaron propiedades que permitieron a *E. coli* explorar nichos ecológicos que de otra manera hubieran sido inalcanzables (16). Si supone que la tasa de transferencia no fue muy diferente para el ancestro común de *Salmonella* y *Escherichia*, es altamente probable encontrar genes que se transfirieron al ancestro común y que han sido conservados en las especies que se originaron de éste.

PLANTEAMIENTO

Los trabajos que han abordado la transferencia horizontal se han centrado sobre todo en los siguientes aspectos: 1) Las especies implicadas en el proceso (1,4,10). 2) Los tipos de genes que son preferentemente transferidos (1,2,3,7). 3) La frecuencia de los eventos de transferencia horizontal (7,16). 4) El papel que juega la transferencia horizontal en el proceso evolutivo de los procariotes en general, y en particular en las bacterias (6,9,10,11). Este último aspecto ha recibido mucha atención en los últimos años pues se ha observado que la transferencia horizontal puede tener un papel importante en el proceso de especiación bacteriana. Hasta ahora ningún estudio ha analizado qué le ocurre a un gen que ha sido transferido en un nuevo genoma.

El objetivo de esta tesis es analizar el proceso evolutivo de los genes transferidos horizontalmente por medio de la determinación del tipo de selección que actúa sobre ellos. Para lograrlo, se desarrolló un modelo que localizó los genes transferidos horizontalmente al ancestro común de los géneros *Escherichia* y *Salmonella* que aún se conservan en dichos géneros.

OBJETIVOS

Objetivo general:

Analizar el proceso evolutivo de los genes adquiridos por transferencia horizontal por el ancestro común de los géneros *Escherichia* y *Salmonella*.

Objetivos particulares:

- 1) Desarrollar un modelo para localizar los genes transferidos.
- 2) Determinar los genes transferidos horizontalmente para el ancestro común de *Escherichia* y *Salmonella*.
- 3) Determinar un grupo control de ortólogos.
- 4) Determinar el tipo de selección que actúa en el grupo control y en los genes transferidos horizontalmente.

ESTRATEGIA EXPERIMENTAL

- 1) Determinación de los genes transferidos.
 - 1.1) Criterio de exclusión aplicado a la búsqueda de homólogos.
 - 1.2) Segundo criterio, que corrobora el primero; incongruencia de los árboles hechos a partir de alineamientos múltiples de los genes transferidos.
- 2) Determinación de genes ortólogos de transferencia vertical, control.
- 3) Determinación del tipo de selección que actúa en los genes transferidos horizontalmente y ortólogos control.
 - 3.1) Valor de la selección promedio.
 - 3.2) Variabilidad de la selección a lo largo de los linajes.
 - 3.3) Variabilidad de la selección en los diferentes sitios de las proteínas.



BIBLIOTECA CENTRAL

MATERIAL Y MÉTODOS

Material

Los materiales utilizados fueron los genomas y proteomas de 67 especies (Apéndice), diversos paquetes de software, programación en el lenguaje Perl y un modelo para la identificación de los genes transferidos horizontalmente. A continuación se describen.

Genomas y proteomas.

Los genomas y proteomas fueron obtenidos del National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/>).

Paquetes de software especializado.

Basic Local Alignment Search Tool (BLAST) . Es el programa más usado para la búsqueda de secuencias en las bases de datos. BLAST realiza comparaciones de secuencias por pares, buscando regiones de similitud local en lugar de hacer alineamientos globales entre el total de las secuencias. BLAST puede realizar miles de comparaciones en cuestión de minutos y en menos de unas cuantas horas la secuencia en cuestión -secuencia sonda- puede ser comparada contra una base de datos entera para encontrar todas las secuencias similares. Uno de los principales usos de BLAST es la identificación de homólogos de cualquier secuencia, para ello fue utilizado en este trabajo. Todos los BLASTs se corrieron con un valor de expectancia de 0.001; el valor de expectancia es la probabilidad de encontrarte al azar un resultado. Todos aquellos "hits" que sus valores de expectancia sean menores a 0.001 estarán incluidos en el BLAST en cuestión. Los archivos con los resultados de las búsqueda de homólogos se denominaron blasts.

CLUSTALW. Es un programa comúnmente utilizado para hacer alineamientos múltiples progresivos. CLUSTALW (el algoritmo que usa) esta basado en el análisis filogenético. Primero una matriz de distancia por pares es generada para todas las secuencias que van a ser alineadas y una guía de árbol es creada usando el algoritmo de neighbor-joining. Luego cada uno de los pares mas relacionados de las secuencias son

alineados uno contra otro. Después cada nuevo alineamiento es analizado para construir un perfil de alineamiento de los pares. Finalmente, los perfiles de alineamiento son alineados uno contra otro -ó contra la secuencias restantes-, hasta que se obtiene el alineamiento total. Esta estrategia de alineamientos es razonable bajo un amplio espectro de condiciones.

Phylogeny Inference Package (PHYLIP versión 3.5). Este es un paquete gratuito para inferir filogénias y llevar a cabo ciertas tareas relacionadas. Actualmente contiene 30 programas, los cuales mediante diferentes algoritmos trabajan diferentes tipos de datos. Los programas que se utilizaron son:

- Protdist: Calcula la distancia entre proteínas con base en el cambio de aminoácidos.
- Fitch: métodos de mínimos cuadrados y FitchMargoliash, para determinar topologías de árboles.
- Consense: La regla de la mayoría y árboles con estricto consenso, cuando se tienen varios árboles y se quiere ver cual es el consenso.
- Drawgram: Dibuja cladogramas y fenogramas.
- Drawtree: Dibuja filogénias sin raíz.

Phylogenetic Analysis by Maximum Likelihood (PAML versión 3.1). PAML es un paquete de programas para análisis filogenético de DNA o proteínas, usando máxima probabilidad. Los posibles usos de los programas son:

- Estimación del largo de las ramas en un árbol filogenético y parámetros en los modelos evolutivos -como la tasa del radio de transición/transversión, el parámetro de la forma de una distribución para el proceso de sustitución-.
- Probar hipótesis concernientes a la evolución de secuencias: como la constancia de las tasas de sustitución; la independencia -para mutar- de los sitios en la secuencias; reloj molecular y la homogeneidad del proceso evolutivo en múltiples genes.
- Cálculos de las tasas de sustitución.
- Reconstrucción de secuencias ancestrales, de nucleótidos o proteínas.
- Simulación de conjuntos de datos - de secuencias de nucleótidos, codones y aminoácidos-.

PAML es útil si uno está interesado en el proceso de evolución de una secuencia. Sus dos principales programas "baseml" y "codeml" implementan modelos sofisticados, que pueden ser usados para construir tests de ratios de probabilidad para las hipótesis evolutivas. El valor resultante (w) de la división de las mutaciones no sinónimas entre las sinónimas, se toma como una medida de la selección que está actuando (anexo II). El programa que se utilizó para determinar el valor promedio de selección fue codeml. Además de obtener el valor de selección promedio codeml puede medir otras dos cosas; variabilidad de la selección a lo largo de los linajes y variabilidad de la selección en los sitios -codones-. Esto lo hace mediante la comparación de modelos, donde cada modelo estipula un tipo particular de condiciones. Las comparaciones de los modelos se realizan por medio del siguiente estadístico: $2(L1-L2)$. Donde $L1$ y $L2$ son el logaritmo de la probabilidad del modelo correspondiente al primer modelo y al segundo modelo. El valor que da el estadístico es comparado con un valor de corte de una tabla de chi cuadrada. Los grados de libertad para estimar el valor de corte son el resultado de la diferencia entre el número de parámetros de los dos modelos; si un modelo tiene tres parámetros y otro cinco, la diferencia es dos; es decir hay dos grados de libertad. Todas las comparaciones se hicieron al 95 % de confiabilidad.

Para ver la variabilidad de la selección en los linajes se utilizaron dos modelos: el primero solo asume un w para todas las ramas del árbol, mientras el segundo asume que cada rama puede tener su propio w . De esta manera el primer modelo asigna un solo valor de w para todos y cada uno de los linajes. Mientras que el segundo puede asignar un valor de w a cada linaje. Se comparan los dos modelos y se ve cual tiene una mayor probabilidad de explicar los datos.

La variabilidad de la selección en los sitios -codones- se verificó utilizando tres modelos: el primero ($M0$) solo asume una categoría -un solo valor de w - para todos los sitios de la proteína; el segundo ($M2$), tiene dos categorías predefinidas $w=0$, $w=1$ - $w=0$ implica que no hay cambio, presión purificadora al máximo y $w=1$ implica que el cambio es neutral, es decir se tienen tantas mutaciones no sinónimas como sinónimas- y la tercera la calcula de los datos, por ende se tiene tres valores de w ; el tercero ($M3$) también utiliza tres categorías pero todas son calculadas de los datos -los tres valores de w se calculan de los datos-. Para las comparaciones se utilizó el estadístico previamente descrito. Las comparaciones se hacen por pares. Primero se compara el primer modelo con el segundo, después el primero con el tercero y por último el segundo con el tercero.

Decir que un modelo obtuvo la mayor probabilidad es decir que ese modelo es el que explica mejor los datos. Si el modelo que tuvo mayor probabilidad fue el primero entonces se concluye que todos los sitios de la proteína tienen el mismo valor de w . Por otra parte si el modelo que obtuvo mayor probabilidad es el segundo se concluye que hay tres tipos de sitios en la proteína; los sitios de la categoría $w=0$ que son sitios bajo presión selectiva purificadora; los sitios de la categoría $w=1$ que son sitios bajo neutralismo; y los sitios de la tercera categoría que tiene un valor igual o mayor a cero. En cambio si el tercer modelo es el que explicó mejor los datos esto implica que hay tres categorías de w calculadas de los datos, que pueden tomar valores de igual o mayor que cero.

Programación en Perl.

Los programas hechos en Perl generalmente sirvieron para concatenar procesos, aunque algunos de ellos tuvieron mayor trascendencia. El programa que determinaba cuáles de los *blasts* cumplían con el proceso de exclusión fue hecho en Perl.

Modelo para la identificación de los genes transferidos

El modelo consiste en localizar de transferencias horizontales antes de la divergencia de *Escherichia* y *Salmonella*; así se tienen dos contextos para el mismo gen transferido, los géneros. Dependiendo del número de especies en que el gen se encuentre presente se tendrían subcontextos. Se supone que el gen incrementa la adecuación, y por eso se conserva.

Se definieron los conjuntos siguientes:

Conjunto EcoSa: se define por todos los genes compartidos por *Escherichia* y *Salmonella*.

Conjunto EcoSa-h: subconjunto de EcoSa, que no contiene a los genes compartidos con los parientes cercanos -posibles genes transferidos-. Este forma el criterio de exclusión.

Conjunto EcoSa+h: subconjunto de EcoSa, que contiene genes compartidos con los parientes cercanos, el control.

La "h" en los conjuntos EcoSa+h y EcoSa-h hace referencia al grupo de especies cercanas que se define como el grupo "hermanos". Una vez localizados los genes, se determinó que tipo de selección operó en ellos y se comparó con la dinámica de los genes transferidos verticalmente.

Las especies de los géneros en cuestión para este análisis fueron: *E. coli*, con tres cepas: K12, O157, O157: H7 EDL933; y dos *Salmonella*: *Salmonella enterica subsp.*

enterica serovar *Typhi*, *Salmonella typhimurium* LT2. Así se tratara de localizar las transferencias horizontales antes de la divergencia de *Escherichia* y *Salmonella*.

Se tienen suficientes genomas cercanos a *Escherichia* y *Salmonella* para recrear el contexto evolutivo que circunscribe a los géneros *Escherichia* y *Salmonella*. Así para el clado de la subdivisión gamma se puede definir el grupo "hermanos" y por ende los conjuntos EcoSa-h y EcoSa+h. Las especies que conforman el grupo "hermanos" (h) son las siguientes: *Yersinia pestis*, *Haemophilus influenzae* Rd, *Pasteurella multocida*, *Buchnera* sp. APS, *Vibrio cholerae*. Estas especies se originan cuatro nodos antes del de *Escherichia* y *Salmonella* en el árbol patrón. Los genes que se encontraron tanto en *Escherichia* y *Salmonella* como los parientes cercanos -grupo "hermanos"- formaron el conjunto EcoSa+h, que fue el grupo control. Los genes que se encontraron en *Escherichia* y *Salmonella* pero que no se encontraron en los parientes cercanos formaron el conjunto EcoSa-h, que es el grupo de los genes transferidos horizontalmente.

El modelo nos permite analizar los genes transferidos sin contar con el donador - esto es de notar, ya que para ver que tipo de selección actúa en cualquier gen se tiene que contar, al menos, con un ortólogo para comparar y detectar los cambios; es decir, ver la mutaciones- y a la vez establecer la dirección de la transferencia, que siempre fue hacia el ancestro de *Escherichia* y *Salmonella*; esto es porque el gen transferido al ancestro común de ambos géneros tiene ortólogos en *Escherichia* y *Salmonella* y no se encuentra presente en los parientes cercanos. De aquí que las comparaciones para ver el tipo de selección se puede hacer entre el ortólogo de cada una de las *Escherichia* y las *Salmonella*.

Metodología

Debido a que el conjunto EcoSa se define por todos aquellos genes que comparten las 5 especies, es suficiente trabajar con una sola especie para la detección de los genes; pues esos genes tienen ortólogos en las cinco especies. Lo primero que se hizo fue buscar los homólogos para todas las proteínas de *E. coli* K12 en una base de datos que contenía los proteomas de los 67 especies, la búsqueda se realizó por medio de BLAST. Los blasts resultantes se analizaron con un programa, hecho en Perl, que determinó cuantos de estos pertenecían al conjunto EcoSa; así como cuales pertenecían al subconjunto EcoSa+h y EcoSa-h -criterio de exclusión-. Todos aquellos blasts en los que estén agrupadas las *Escherichia* y *Salmonella* -en los primeros cinco lugares- e inmediatamente después cualquier otra especie que no sea las cinco especies del grupo

"hermanos" son blasts que cumplieron el criterio de exclusión; es decir la secuencia en cuestión sería un gen transferido horizontalmente; estos blasts definieron el conjunto EcoSa-h. El siguiente paso fue corroborar la ortología entre las *Escherichia* y *Salmonella* para los genes de EcoSa-h; esto se hizo mediante mejores "hits" bidireccionales (Bidirectional best hits). Esto es si la proteína "a" fue el mejor "hit" de la proteína "b" en el genoma "A", la proteína "b" también tiene que ser el mejor "hit" de la proteína "a" en el genoma "B". Después, para confirmar que el subconjunto EcoSa-h contenía realmente genes transferidos horizontalmente, se hicieron los árboles de estos genes y se compararon con el árbol patrón - árbol hecho con la proteína del factor de elongación Tu (EF-Tu) de *E. coli* K12 y con sus ortólogos para las otras 66 especies; este árbol es congruente con el árbol de la vida.- (Anexo I). Todos los árboles fueron hechos con métodos de distancia, con PHYLIP. Si realmente son genes transferidos deben presentar incongruencias con el árbol patrón (Anexo I). Así el primer criterio para definir el conjunto de genes transferidos es el criterio de exclusión, y después dicho conjunto es corroborado por el segundo criterio -la incongruencia de los árboles de los genes transferidos con el árbol patrón-.

Para analizar la posible ubicación, el clado, de una especie originada del donador - el clado en el que se podría encontrar esta especie es el mismo en que se encontraba el donador-, se observaron las topologías de los árboles de los genes transferidos. Pero sobre todo, se observó si la proteína de la supuesta especie originada del donador tenía uno de sus mejores "hits", al hacerle un BLAST contra toda la base de datos, con alguna proteína de *Escherichia* o *Salmonella*.

Una vez que se tuvo el conjunto de los genes transferidos, se procedió a observar el tipo de selección que actuaba sobre ellos. Para eso se diseñó un programa en Perl que alinea los codones de las secuencias, de las tres *Escherichia* y las dos *Salmonella*, tomando como referente los alineamientos de las proteínas para los que codifican dichos codones. Por lo tanto, quedaron alineados sólo codones que comparten común ancestría.

Ya que se tuvieron los alineamientos de los codones, se analizaron con PAML para ver que tipo de selección actuó. Esto se realizó con "codeml", un programa de PAML, el cual se utiliza para analizar codones.

RESULTADOS

Localización de genes

De acuerdo a los criterios utilizados fueron nueve los genes que se identificaron como transferidos horizontalmente (Tabla 1). Estos nueve genes cumplieron el criterio de exclusión, en los blasts, y cuando se realizaron los respectivos árboles de los genes los árboles presentaron topologías que diferían considerablemente del árbol patrón. Solo dos de ellos, gen 7 y gen 4, tienen función asignada, cinco tienen una función probable y dos más no se tiene ningún indicio de su función, gen 8 y gen 6. Los ortólogos de estos genes, en las otras especies, se encuentran en la Tabla 2.

Tabla 1

Genes transferidos horizontalmente (para *E. coli* K12)

Gen	Identificador	Función
1	Gi 16130968	Probable aminotransferasa de la ornitina
2	Gi 16130627	Probable proteína hierro-azufre de la hidrogenasa 3 (parte del complejo FHL)
3	Gi 16130444	Enzima del peptidoglicano putativa
4	Gi 16130383	Utilización de etanolamina; homólogo de la acetil/butiril P transferasa de Salmonella
5	Gi 16129998	Transferasa putativa
6	Gi 16129797	Orf, Proteína hipotética
7	Gi 16129791	Proteína fosfatasa 1 (modula fosfoproteínas, señalizando plegamientos incorrectos)
8	Gi 16129291	Orf, Proteína hipotética
9	Gi 16128289	Oxidoreductasa putativa

Tabla 2

Ortólogos de los genes transferidos

Gen	Identificador O157:H7	Identificador EDL933	Identificador LT2	Identificador TYPHI
1	15833209	15803614	16766517	16761991
2	15832830	15803237	16766154	16761625
3	15832639	15803046	16765851	16761445
4	15832574	15802980	16765786	16761383
5	15832117	15802538	16765444	16761044
6	15831808	15802257	16765220	16760829
7	15831802	15802251	16765194	16760741
8	15831166	15801837	16765007	16760224
9	15829596	15800018	16763941	16759530

Abreviaturas

E. coli O157:H7(O157:H7); E. coli O157:H7 EDL933(EDL933)

S. enterica subsp. enterica serovar Thyphi(TYPHI); S. typhimurium LT2(LT2)

De los nueve genes solo los genes 9, 3, 2 y 1, permiten la identificación del clado al que podría pertenecer el posible donador. Pues al ver las topologías de los árboles de estos genes parece que el clado que forman las *Escherichia* y *Salmonella* hubiese sido movido a otro clado, conservándose en general el orden de los clados. Cuando se hizo el BLAST para la proteína de la supuesta especie originada del donador, los ortólogos –de la Tabla 2- del gene transferido fueron de sus mejores hits. Entre los clados de los posibles donadores encontramos incluso el caso de un clado donador perteneciente a las arqueas. Los demás clados donadores se ubican dentro de las bacterias, con representantes variados (Tabla 3).

Tabla 3

Clados de los posibles donadores

Gen	Clado del donador
9	Grupo de Bacillus/Staphylococcus(Gpo. Bacillus/Clostridium,Firmicutes)
2	Thermococcales(Euryarchaeota, Arqueas)
3	Grupo de las Xanthomonas(subdivision gamma, Proteobacteria)
1	Grupo de Ralstonia (subdivision Beta, Proteobacteria)

El grupo control, el conjunto EcoSa+h, fue de 208 genes. Este grupo esta compuesto por 70 genes informacionales, 41 genes que son proteínas hipotéticas o tienen funciones putativas, y el resto son 97 genes operacionales.

Valor promedio de la selección.

Para este, y posteriores, análisis fueron eliminadas las secuencias de *E. coli* O157:H7 EDL933, pues se vió que era exactamente iguales que las de *E.coli* O157:H7; de tal forma que para el análisis de selección solo se trabajo con dos *Escherichia* y dos *Salmonella*.

Para ninguno de los genes se obtuvo un valor mayor de 1, no llegó incluso a 0.5. El valor más alto fue 0.11 y el más bajo de 0.02; dando un rango 0.0831 (Tabla 4). Los otros valores no se cargan ni al valor más alto ni al más bajo. La media de estos valores fue 0.0684 con una varianza de 0.00097. Un aspecto que hay que tomar en cuenta es que esta determinación de w se realiza a partir de un promedio de todos los sitios; de tal manera que si hay mayor proporción de sitios donde la división de la mutaciones no sinónimas entre las mutaciones sinónimas (dn/ds) está por debajo de 1, que de sitios con dn/ds arriba de uno, el radio del linaje w será por debajo de 1. De tal manera que este valor de w solo nos indica que pasa en promedio en la proteína.

Tabla 4

Medias de w y sus varianzas

Gen	Media	Varianza
1	0.0278666666666667	0.0000845788888888889
2	0.0446333333333333	0.000549138888888889
3	0.104	0.002362023333333333
4	0.0925	0.00877026666666667
5	0.0905833333333333	0.0142551780555556
6	0.0302	0.000196216666666667
7	0.1111833333333333	0.0087836347222222
8	0.0388833333333333	0.00351247138888889
9	0.07585	0.00322443916666667

Para el grupo control tampoco se alcanzó un valor de w igual a 1. El valor más alto fue de 0.733 y el valor más bajo fue de 0; dando un rango de 0.73. La media de dichos valores dió un resultado de 0.1078 y una varianza de 0.00764; si sumamos y restamos dos desviaciones estándar a la media, se ve que el 95 por ciento de los datos se encuentra entre 0 y 0.2826. Como se aprecia la mayoría de los datos esta muy por debajo un valor de w de 0.5.

Variabilidad de la selección en los linajes.

Posteriormente se analizó que tanto variaba el valor de w en los linajes, esto se hizo comparando los modelos M0 y M1 (Anexo II). Esto se hace mediante el estadístico antes descrito –en material-, el cual compara dos modelos, cualesquiera, y ve cual explica mejor los datos; cual tiene una probabilidad más alta de explicar lo datos. Cuando se compararon los dos modelos para los nueve genes solo los genes 8 y 3 –que fue un 22.2% de los nueve genes- fueron mejor explicados por el modelo que asume un radio para cada rama. Para los otros siete genes no hay casi variación en los valores de w de las diferentes ramas (Tabla 5). Cuando se analizó que tanto variaba el valor de w a lo largo de los linajes en el grupo control, de los 208 genes, solo 15 fueron variables lo cual es una proporción muy pequeña -7.2 %-.

Tabla 5

Variabilidad de w entre los linajes

Comparación del modelo de un solo w para todas las ramas y del modelo de un w para cada rama

Grados de libertad 4 .Valor de corte 9.48777 (al 95 % de confiabilidad)

Gen	Valor de chiquadrada	Variabilidad entre linajes
1	1.66663799999969	No
2	5.08491000000004	No
3	10.30602200000007	Si
4	8.42245999999977	No
5	4.83602199999996	No
6	0.736113999999816	No
7	5.40550399999984	No
8	11.517108	Si
9	2.97135400000025	No



Variabilidad de la selección en los sitios.

Para medir la variabilidad de la selección en los diferentes sitios -codones-, de los genes, también se hizo una comparación de diferentes modelos para ver cual explicaba mejor los datos. Las comparaciones fueron, al igual que antes, en pares: 1) M0 vs M2, 2) M0 vs M3 y 3) M2 vs M3 (Anexo II).

En la primera comparación, M0 vs M2, cinco fueron los genes mejor explicados por M2, 9, 8, 7, 4, y 3 (Tabla 6.1). Cuando se hizo la segunda comparación, M0 vs M3, fueron exactamente los mismos cinco genes los que ahora fueron mejor explicados por M3 (Tabla 6.2). En la última comparación, M2 vs M3, todos lo genes fueron mejor explicados por M2 (Tabla 6.3). Esto indica que al parecer solo cinco genes, los antes mencionados, tienen diferentes valores de selección (w) a lo largo de la proteína; mientras que los restantes parecen ser muy bien explicados por un solo valor de selección, que es el valor promedio de w calculado para toda la proteína (Tabla 4). En todos los casos, salvo el gen 7, la tercer categoría de M2 no pasó de un valor de w de 0.5; el valor más alto fue de 0.56541 y el más bajo fue de 0.01105, dando un rango de 0.55436 (Tabla 7.1). Al ver la proporción de las categorías (p) para M2, es posible observar que la mayoría de los datos son explicados por la categoría 0 y la tercer categoría, la calculada por los datos. Así la mayoría de los datos quedan entre dos valores de $w=0$ y $w=0.5$ (Tabla 7.1). Una parte mínima de los datos, más del 5%, de cada uno de los genes 8, 4 y 3 cayó en la categoría 1 -esta supone neutralismo-. La proporción de datos de la categoría 1 fue de 6.1% para el gen 8, 8% para el gen 4 y de 11.2 % para el gen 3 (Tabla 7.1). Solo para dos genes hubo sitios con presión selectiva positiva: 11 sitios en el gen 4, que representan un 1% de los datos del gen; 104 sitios en el gen 3, que representan un 4.5% de lo datos del gen.

Tablas 6.1, 6.2 y 6.3
Variabilidad de w en los sitios de las proteínas

6.1 Comparación M0 vs M2
Valor de corte 5.9915

6.2 Comparación M0 vs M3
Valor de corte 9.4877

Gen	Valor	Mejor explicado Por M2	Gen	Valor	Mejor explicado por M3
1	1.147022000000	No	1	1.146847999999	No
2	5.960166000000	No	2	5.959703999999	No
3	128.5214600000	Si	3	128.7594860000	Si
4	29.85142400000	Si	4	30.57413999999	Si
5	3.371853999999	No	5	3.475619999999	No
6	2.095463999999	No	6	2.095305999999	No
7	46.31880600000	Si	7	46.31382600000	Si
8	37.39044800000	Si	8	37.68105199999	Si
9	47.62403199999	Si	9	47.64927600000	Si

6.3 Comparación M2 vs M3
Valor de corte 5.9915

Gen	Valor	Mejor explicado por M3
3	0.238025999999081	No
4	0.722715999999309	No
5	0.103765999999951	No
8	0.290603999999803	No
9	0.025244000000384	No

Tablas 7.1 y 7.2

Categorías y frecuencias para el segundo(M2) y el tercer modelo(M3)

7.1 Proporción de los datos para cada categoría del segundo modelo

Gen	W	Proporción	W	Proporción	W	Proporción
3	0	0.00000	1	0.15721	0.02015	0.84279
4	0	0.00000	1	0.09061	0.02018	0.90939
5	0	0.00000	1	0.02909	0.03539	0.97091
8	0	0.41176	1	0.06106	0.01105	0.52718
9	0	0.60789	1	0.03639	0.18578	0.35573

7.2 Proporción de los datos para cada categoría del tercer modelo

Gen	W	Proporción	W	Proporción	W	Proporción
3	0.02330	0.15132	0.02330	0.70417	1.17459	0.14451
4	0.02953	0.46206	0.02953	0.47931	1.71398	0.05863
5	0.03734	0.45087	0.03734	0.52675	1.51631	0.02238
8	0.00010	0.36525	0.01063	0.56599	0.65025	0.06876
9	0.00010	0.61173	0.19306	0.35908	1.23169	0.02919

El grupo control tuvo 155 genes mejor explicados por M2 en la primera comparación, M0 vs M2. Para la segunda comparación, M0 vs M3, 149 genes fueron mejor explicados por M3. En la tercera comparación, M2 vs M3, solo 27 genes son mejor explicados por M3. Dentro del grupo de 53 genes que son mejor explicados por M0 se obtuvo una media de los valores de w de 0.126 y una varianza de 0.0182; el valor más alto fue 0.5854 y el más bajo 0.0085. Entonces, el 95 % de los datos se encuentran entre los valores de 0.3958 y 0. En el segundo grupo -que comprende los 155 genes mejor explicados por M2 en la comparación 1 menos los 27 genes que fueron mejor explicados por M3 en la tercera comparación -, el mejor explicado por M2, se observó que la categoría de w igual a 0 tiene una proporción de 0.82050 -82 %- , mientras que la categoría de w igual a 1 tiene una proporción de 0.03820 -casi el 4 %- y la categoría tres, calculada de los datos, mostró una proporción de 0.149040 -15 %- . Al observar bajo qué valores de w la categoría tres explicaba más datos, se obtuvo lo siguiente: el rango de 0 a 0.5 explicó hasta 0.13098; el siguiente rango de 0.5 a 1 explicó hasta 0.00667 y el rango arriba de 1 explicó hasta 0.01139. Si se suma lo que explicó cada rango da la proporción 0.149040 de la categoría tres. El 95 % de los datos de este grupo cae en un rango de 0 a 0.5 -si se suma la proporción de la categoría 0, 0.82050, y la proporción del primer rango

de la categoría tres, 0.13098, resulta una proporción con un poco más que el 95 % de los datos- y se ve que un 4 % de los datos se comporta de manera neutral y poco más de un 1 % ,0.01139, podría estar sufriendo presión selectiva positiva. Para el tercer grupo, el mejor explicado por M3, se procedió de manera similar, se dividieron todos los valores de w en tres rangos y se observó cuanta proporción de los datos era explicada por cada rango. El rango que fue de 0 a 0.7 tuvo una proporción de 0.939098 -un 94 %- . En el rango de 0.7 a 1.2 se obtuvo una proporción de 0.039380 -un 4 %- mientras que para el último rango, arriba de 1.2, la proporción fue de 0.02152 -2 %- . Para este grupo el 95 % de los datos está entre 0 y 0.9; un aspecto a notar es que en este grupo hay un 2 % que estaría sujeto a presión selectiva positiva. Aunque el segundo rango, 0.7 a 1.2 -que explica un casi un 4%- no es estrictamente un proceso neutral, si se aproxima a este. Puede parecer arbitrario que el rango de 0.7 a 1.2 sugiera un proceso muy parecido al neutral, pero no lo es tanto; son realmente pocos los sitios que dan exactamente un valor de $w = 1$, incluso para la categoría 1 de M2 no toda la proporción que explica tiene un valor w de 1. Lo que realmente se tiene son una serie de sitios que su valor está más próximo a uno que a cualquiera de las otras dos categorías.

DISCUSIÓN

Los genes transferidos horizontalmente localizados en este estudio fueron traspasados hace más de 100 millones de años. El hecho de encontrarse actualmente, en los genomas, de las especies, de los géneros *Escherichia* y *Salmonella*, implica que dichos genes incrementaron la adecuación del ancestro común de ambos géneros y aun ahora son importantes en la adecuación de estos géneros.

Al observar los cladogramas donadores (Tabla 3) uno ve que hay distancias filogenéticas muy variadas, desde un clado en la misma subdivisión, como sería el grupo de las *Xanthomonas* -lo cual representaría unos cuantos de cientos de millones de años-, hasta el caso del clado de los *Thermococcales*, el cual se encuentra en el reino *Euryarchaeota*, en el imperio de las arqueas -lo cual representaría la magnitud de miles de millones de años-. Los otros dos casos tienen distancias intermedias como son la subdivisión beta -dentro de la propia división de *Escherichia* y *Salmonella*, la de las *Proteobacterias*- y el grupo de *Bacillus/Staphylococcus* -en otra división, la de los *Firmicutes*-. Esto corrobora lo que han expuesto otros trabajos (4,10), que la transferencia se puede llevar a cabo incluso entre grandes distancias filogenéticas. Para el caso de los otros cinco genes en los que no se identificó un clado donador, la respuesta más plausible parecería ser que hasta el momento no se han secuenciado las especies de dichos cladogramas -o a lo mejor no en número suficiente; es decir, hay varios cladogramas que solo están representados por una especie-.

Como se pudo observar tanto en los genes transferidos horizontalmente como en el grupo control los valores promedio de la selección quedaron muy por debajo de uno. En los genes transferidos el valor más alto de w fue de 0.111, con una media de los valores de 0.0684 y con una varianza 0.00097. Cuando se analiza el grupo control se aprecia que, aunque el valor más alto es de 0.733, la media de los valores es 0.1078 y tiene una varianza de 0.00764; quedando el 95 % de los datos por debajo de 0.29. De aquí se concluye que en los dos grupos, el promedio de w es sumamente bajo, implicando selección purificadora; al parecer en los genes transferidos esta es de mayor intensidad. Si se comparan, los valores obtenidos, con el valor de $w = 0.045$ que obtuvieron Jordan I.

K. et al, para los genes esenciales, (21) se ve que son muy parecidos. Entonces cuando hay una transferencia exitosa, que incrementó la adecuación del organismo, la selección actúa de tal manera que permite muy pocos cambios significativos -no sinónimos-, dejando que la estructura de la proteína, y por ende su función, permanezca casi intacta. Teniendo una intensidad, la selección purificadora, muy parecida a la de los genes esenciales. Esto parece ser la norma en el promedio de la proteína.

Al observar la variabilidad de la selección en los linajes, se aprecia que esta se ve incrementada en los genes transferidos horizontalmente; donde un 22.2% de los genes fue variable comparado con un 7.2 % de genes para el grupo control. Así para el grupo de los genes de descendencia vertical la selección, el valor de w , en general parece ser más constante en el proceso de microevolución. Aunque si se analizan los valores medios y las varianzas de los genes transferidos se verá que esta variación es mínima. Para los dos genes que presentaron variación de w a lo largo del linaje, gen 3 y gen 8, se ve que 95 % de los datos, para el primero quedaría entre 0.152 y 0.055, y para el segundo entre 0.04231 y 0.0352. Estos dos rangos se ubican perfectamente como una selección purificadora intensa; mas aún solo el valor de corte superior, 0.152, para el gen 3 se sale del 95 % del total de los valores promedio de w de los genes transferidos, pero dicho valor se encuentra perfectamente dentro del 95 % de los valores promedio de w para el grupo control. Entonces, aunque algunos genes transferidos varíen sus valores de w , los mantienen en valores muy bajos que no llegan ni al 0.16.

La variabilidad de la selección en los sitios parece corroborar que para los genes transferidos el principal componente evolutivo es la selección purificadora. Para todos los genes transferidos el 80 % de sus sitios caen en un rango de w que va de 0 a 0.5. A excepción de los genes 8, 4 y 3 todos los demás genes transferidos tienen el 95 % de los sitios en el rango antes mencionado. Para estos tres se encontró que algunos sitios obedecen al neutralismo, 6.1 % para el gen 8, 8 % para el gen 4 y 11.2 % para el gen 3; y que el gen 4 tanto como el gen 3 tienen 1% y 4.5% respectivamente de sitios con presión selectiva positiva. Estos dos genes fueron además los dos únicos genes que presentaron un valor de w variable entre de los linajes. Entonces al parecer en los genes transferidos tenemos dos subgrupos: uno en el que 95 % de los sitios sufren presión selectiva negativa - éste representa 45 % de los genes transferidos -, y el segundo en el que aunque la presión selectiva negativa es el principal tipo de selección, dando cuenta de más del 80 % de los sitios, habría sitios que proceden de manera neutral y que habría

algunos de estos genes que tienen incluso sitios con presión selectiva positiva –éste es el 55 % de los genes transferidos-. Este subgrupo tuvo genes que su valor de w fue variable entre los linajes, aunque dicha variabilidad queda en un rango pequeño, que se sigue ubicando como presión selectiva negativa.

El grupo control se puede dividir en tres subgrupos cuando se observa la variabilidad en la selección en los sitios. El primero, el mejor explicado por M0, fue aquel en que el 95 % de los datos estuvo entre los valores de w de 0 y 0.3958; incluso el total de los datos no pasa de 0.6. En este subgrupo, de 53 genes, el principal y casi único elemento es la presión selectiva negativa de alta intensidad. Para el segundo subgrupo, el que fue mejor explicado por M2, el 95 % de los datos está en el rango de valores de w que va de 0 a 0.5. Casi un 4 % de los sitios se ubica en la categoría de w igual 1 y poco más del 1 % que tiene valores más altos que uno. Así para este subgrupo, de 129 genes, aunque la presión selectiva es el principal componente –pero debe tenerse en cuenta que es de menor intensidad que la del primer grupo, compárense los rangos-, se encuentra que alrededor de 4 % de los sitios tiene un proceso parecido al neutral y un 1 % parece tener presión selectiva positiva. Para el tercer subgrupo, el mejor explicado por M3, el 95 % de los datos se encontró en un rango de valores de w que fue de 0 a 0.9. Se encontró que casi el 4 % se encuentra en el rango de 0.7 a 1.2 y que un 2 % se pasa el valor de 1.2. Por ende, para este subgrupo de 27 genes la presión selectiva negativa es el principal componente, 94 % de los sitios, aunque es más bien una presión laxa de un amplio rango. El segundo elemento son sitios con procesos parecidos al neutral con casi un 4 % y hay un 2 % que son sitios con presión selectiva positiva.

Al comparar los genes transferidos horizontalmente con los genes del grupo control –para la variabilidad de la selección en los sitios-, se ve que en ambos su primer subgrupo presenta a la selección purificadora, de alta intensidad, como primera y única opción –estos subgrupos fueron los mejor explicados por M0-. El segundo subgrupo de los genes transferidos se homologaría sobre todo con el tercer subgrupo de los genes control; en ambos aunque la presión selectiva negativa es el factor principal, es una presión más laxa y con menor intensidad, estando el neutralismo en segundo lugar y con sitios con presión selectiva positiva –neutralismo y presión positiva explicaron más del 5 % de los datos-. El segundo subgrupo de los genes control se parece a lo anteriores en que el también presenta neutralismo en segundo lugar y presión selectiva en tercero. Pero estos dos procesos juntos no llegan a explicar el 5% de los datos, esta es la primera diferencia.

La segunda diferencia es que su presión selectiva negativa no es tan laxa, pues esta en un rango de 0 a 0.5 donde caen el 95 % de los datos.

Cuando se tiene una transferencia exitosa, en el sentido de incrementar la adecuación del organismo, parece que el principal componente es la presión purificadora. Esto asegura que la función que se introdujo, y que repercutió para bien de la especie, se mantenga casi intacta. Pero por otro lado no hay que olvidar que algunos sitios presentaron neutralismo y unos pocos presión selectiva positiva; esto por su parte iría amoldando el gen al genoma en que se encuentra. De tal manera que el gen no es inmutable, pero ese cambio nunca es a costa de poner en riesgo la función; así, son realmente pocos los sitios que cambian. Esto no era lo que se esperaba *a priori*, pues como el gen vino de otro genoma, se espera que la presión selectiva positiva - y en el peor de los casos, el neutralismo- jugara un papel más preponderante; ya que esta hubiera amoldado el gen a su nuevo genoma. Pero al parecer este proceso de amoldamiento está sujeto a que se mantenga la función; de tal manera que dicho proceso es permitido en tanto que no se afecte la eficacia de la función.

CONCLUSIONES

Se elaboró y aplicó un modelo que identificó los genes que fueron transferidos horizontalmente al ancestro común de *Escherichia* y *Salmonella*.

Se localizaron nueve genes transferidos horizontalmente antes de la divergencia de *Escherichia* y *Salmonella* (Tabla 1). Solo de cuatro fue posible establecer el clado del posible donante.

Para el grupo control, se localizaron 208 genes, que son los genes de descendencia vertical. En este grupo encontramos tanto genes operacionales (97 genes) como informacionales (70 genes) y genes que no tiene función asignada (41 genes).

Al comparar la dinámica de los dos grupos parece ser que para los genes transferidos horizontalmente así como para el grupo control la selección negativa es el componente principal en el promedio de la proteína. Sin embargo, ambos grupos tuvieron subgrupos de genes en los cuales hay algunos sitios que presentaron neutralismo y presión selectiva positiva.

ANEXO I

DetECCIÓN DE LAS TRANSFERENCIAS.

Una variedad de métodos han sido desarrollados para inferir la ocurrencia de transferencia horizontal. Este anexo solo se referirá a los métodos de la biología computacional que son más usados y no a los de orden experimental.

Patrones de mejor puntaje para diferentes especies. Una alternativa común para inferir transferencia horizontal es usando técnicas de búsqueda de similitud para determinar el mejor puntaje para cada gen de un genoma. La transferencia horizontal es frecuentemente invocada para aquellos genes que tienen el mejor puntaje con especies supuestamente distantes, evolutivamente hablando (9,11). Para los patrones de mejor puntaje se corrieron BLASTs y se aplicó el criterio de exclusión antes mencionado. De esta manera los genes homólogos más cercanos, con puntajes más altos, a las *E. coli* y *Salmonella* que pertenecieron a especies que no son los parientes cercanos de ambas son los posibles genes transferidos horizontalmente.

Comparaciones de árboles filogenéticos de diferentes genes. Un método que ha sido usado para determinar si ha ocurrido transferencia horizontal consiste en inferir árboles evolutivos para muchos genes en muchos genomas. Primero se determinan los ortólogos del gen de interés en las diferentes especies. Segundo se realiza el árbol. Tercero se compara la topología del árbol con la topología del árbol base -árbol de la subunidad pequeña ribosomal-. La transferencia horizontal debería causar que genes transferidos horizontalmente difieran en su topología con la topología del árbol base. Es importante reconocer que los métodos de reconstrucción filogenética no son perfectos. Sin embargo, los métodos de reconstrucción filogenética son la única manera de inferir eventos históricos de los genes de manera confiable (9,11).

El único método que presenta un contexto histórico de los genes en cuestión es la comparación de árboles filogenéticos; es hasta ahora el método más robusto para la inferencia de transferencia horizontal. De cualquier manera en este trabajo también se utilizó el método de patrones de mejor puntaje. Cuando consideramos estos métodos, uno tiene que tener en mente que pruebas directas de transferencia horizontal quizá sean

inaccesibles, por la simple razón que no hay otro trazo de estos eventos evolutivos que aquel que se infiere por la comparación de los genomas actuales. Por lo tanto, todas las indicaciones de transferencia horizontal necesariamente permanecen en probabilidad, y el punto de usar diferentes métodos es para maximizar la probabilidad de que estos eventos sean identificados correctamente (10).



ANEXO II

Midiendo la selección (w).

Tradicionalmente las tasas de sustituciones sinónimas y no sinónimas son definidas comparando dos secuencias de DNA, con d_s y d_n como los números de sustituciones sinónimas y no sinónimas por sitio respectivamente. El radio $w = d_n/d_s$ mide la diferencia entre las dos tasas. Si un cambio de aminoácido es neutral, este será fijado a la misma tasa que las mutaciones sinónimas, con $w=1$. Si el cambio de aminoácido es deletéreo, el proceso de selección purificadora reducirá su tasa de fijación, siendo entonces $w < 1$. Solo cuando el cambio de aminoácido ofrece una ventaja selectiva es fijado a una tasa mayor que la tasa de mutaciones sinónimas, con $w > 1$. Por tanto, un w más grande que uno es una evidencia convincente de selección diversificadora (14). Con w se pretende saber que tipo de selección es la que opera en los genes transferidos horizontalmente.

Dos clases de métodos han sido sugeridos para estimar d_n y d_s entre dos secuencias. La primera clase incluye mas de una docena de métodos intuitivos desarrollados desde los comienzos de los ochentas. Estos métodos involucran los siguientes pasos:

- a) Contar los sitios sinónimos S y no sinónimos N en las dos secuencias.
- b) Contar las diferencias sinónimas y no sinónimas entre las dos secuencias, y corregir para cambios múltiples.

La mayoría de estos métodos hacen supuestos simples acerca del proceso de sustitución de aminoácidos y también hacen tratamientos *ad hoc* de los datos, que no siempre son justificados. Por eso se refieren a esos métodos como métodos aproximados. La segunda clase de métodos es la de máxima probabilidad, basados en modelos explícitos del proceso de sustitución de codones. Los parámetros de los modelos -tiempo de divergencia de las secuencias t , tasa del radio de transiciones/transversiones k , etc.- son estimados de los datos por máxima probabilidad, y son usados para calcular d_n y d_s de acuerdo a sus definiciones.

Dos problemas afectan de manera contundente la estimación de w . Uno es el sesgo en el uso de codones. El otro es el sesgo en el ratio transición/transversión. Ignorando el sesgo en el uso de codones lleva a un sobrestimado de S , un subestimación de ds y, por ende, una sobrestimación de w . Si se ignora el sesgo en el ratio transición/transversión se tiene un subestimación de S , sobrestimación de ds y, ergo, una subestimación de w .

Debido a que los métodos de máxima probabilidad lidian con estos problemas de manera más directa -debido a los parámetros en los modelos-, de entre estos se escogió el paquete de programas Phylogenetic Analysis by Maximum Likelihood (PALM) (15).

APÉNDICE

Genomas utilizados

Imperio	Reino	Sección	Especie
Arquea	Crenarchaeota	Desulfurococales	<i>Aeropyrum pernix</i>
Bacteria	Eubacteria	Aquificales	<i>Agrobacterium tumefaciens</i>
Bacteria	Eubacteria	Aquificales	<i>Agrobacterium tumefaciens str. C58 Dupont</i>
Arquea	Euryarcheota	Archaeoglobales	<i>Aquifex aeolicus</i>
Bacteria	Firmicutes	Grupo Bacillus/ Clostridium	<i>Archaeoglobus fulgidus</i>
Bacteria	Firmicutes	Grupo Bacillus/ Clostridium	<i>Bacillus halodurans</i>
Bacteria	Spirochaetales	Spirochaetaceae	<i>Bacillus subtilis</i>
Bacteria	Proteobacteria	Subdivisión alfa	<i>Borrelia burgdorferi</i>
Bacteria	Proteobacteria	Subdivisión gama	<i>Brucella melitensis</i>
Bacteria	Proteobacteria	Subdivisión epsilon	<i>Buchnera sp. APS</i>
Bacteria	Proteobacteria	Subdivisión alfa	<i>Campylobacter jejuni</i>
Bacteria	Chlamydiales	Chlamydiaceae	<i>Caulobacter crescentus</i>
Bacteria	Chlamydiales	Chlamydiaceae	<i>Chlamydia muridarum</i>
Bacteria	Chlamydiales	Chlamydiaceae	<i>Chlamydia trachomatis</i>
Bacteria	Chlamydiales	Chlamydiaceae	<i>Chlamydophila pneumoniae AR39</i>
Bacteria	Chlamydiales	Chlamydiaceae	<i>Chlamydophila pneumoniae CWL029</i>
Bacteria	Firmicutes	Grupo Bacillus/ Clostridium	<i>Chlamydophila pneumoniae J138</i>
Bacteria	Grupo Thermus/ Deinococcus	Deinococcales	<i>Clostridium acetobutylicum</i>
Bacteria	Proteobacteria	Subdivisión gama	<i>Deinococcus radiodurans</i>
Bacteria	Proteobacteria	Subdivisión gama	<i>Escherichia coli K12</i>
Bacteria	Proteobacteria	Subdivisión gama	<i>Escherichia coli O157:H7</i>
Bacteria	Proteobacteria	Subdivisión gama	<i>Escherichia coli O157:H7 EDL933</i>

Arquea	Euryarchaeota	Halobacteriales	<i>Haemophilus influenzae</i> Rd
Bacteria	Proteobacteria	Subdivisión epsilon	<i>Halobacterium</i> sp. NRC-1
Bacteria	Proteobacteria	Subdivisión epsilon	<i>Helicobacter pylori</i> 26695
Bacteria	Firmicutes	Grupo Bacillus/ Clostridium	<i>Helicobacter pylori</i> J99
Bacteria	Firmicutes	Grupo Bacillus/ Clostridium	<i>Lactococcus lactis</i> subsp. <i>lactis</i>
Bacteria	Firmicutes	Grupo Bacillus/ Clostridium	<i>Listeria innocua</i> ,
Bacteria	Proteobacteria	Subdivisión alfa	<i>Listeria monocytogenes</i> EGD-e
Arquea	Euryarchaeota	Methanococcales	<i>Mesorhizobium loti</i>
Arquea	Euryarchaeota	Methanobacteriales	<i>Methanococcus jannaschii</i>
Bacteria	Firmicutes	Actinobacteria	<i>Methanothermobacter thermautotrophicum</i>
Bacteria	Firmicutes	Actinobacteria	<i>Mycobacterium leprae</i>
Bacteria	Firmicutes	Actinobacteria	<i>Mycobacterium tuberculosis</i> CDC1551
Bacteria	Firmicutes	Grupo Bacillus/ Clostridium	<i>Mycobacterium tuberculosis</i> H37Rv
Bacteria	Firmicutes	Grupo Bacillus/ Clostridium	<i>Mycoplasma genitalium</i>
Bacteria	Firmicutes	Grupo Bacillus/ Clostridium	<i>Mycoplasma pneumoniae</i>
Bacteria	Proteobacteria	Subdivisión beta	<i>Mycoplasma pulmonis</i>
Bacteria	Proteobacteria	Subdivisión beta	<i>Neisseria meningitidis</i> MC58
Bacteria	Cyanobacteria	Nostocales	<i>Neisseria meningitidis</i> Z2491
Bacteria	Proteobacteria	Subdivisión gama	<i>Nostoc</i> sp. PCC 7120
Bacteria	Proteobacteria	Subdivisión gama	<i>Pasteurella multocida</i>
Arquea	Euryarchaeota	Thermococcales	<i>Pseudomonas aeruginosa</i>
Arquea	Euryarchaeota	Thermococcales	<i>Pyrococcus abyssi</i>
Bacteria	Proteobacteria	Subdivisión beta	<i>Pyrococcus horikoshii</i>
Bacteria	Proteobacteria	Subdivisión alfa	<i>Ralstonia solanacearum</i>
Bacteria	Proteobacteria	Subdivisión alfa	<i>Rickettsia prowazekii</i>
Bacteria	Proteobacteria	Subdivisión gama	<i>Rickettsia conorii</i>

Bacteria	Proteobacteria	Subdivisión gama	<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar <i>Typhi</i>
Bacteria	Proteobacteria	Subdivisión alfa	<i>Salmonella typhimurium</i> LT2
Bacteria	Firmicutes	Grupo Bacillus/ Clostridium	<i>Sinorhizobium meliloti</i>
Bacteria	Firmicutes	Grupo Bacillus/ Clostridium	<i>Staphylococcus aureus</i> subsp. <i>aureus</i> Mu50
Bacteria	Firmicutes	Grupo Bacillus/ Clostridium	<i>Staphylococcus aureus</i> subsp. <i>aureus</i> N315
Bacteria	Firmicutes	Grupo Bacillus/ Clostridium	<i>Streptococcus pneumoniae</i> R6
Bacteria	Firmicutes	Grupo Bacillus/ Clostridium	<i>Streptococcus pneumoniae</i> TIGR4
Arquea	Crenarchaeota	Sulfolobales	<i>Streptococcus pyogenes</i>
Arquea	Crenarchaeota	Sulfolobales	<i>Sulfolobus solfataricus</i>
Bacteria	Cyanobacteria	Chroococcales	<i>Sulfolobus tokodaii</i>
Arquea	Euryarchaeota	Thermoplas Males	<i>Synechocystis</i> sp. PCC 6803
Arquea	Euryarchaeota	Thermoplas Males	<i>Thermoplasma acidophilum</i>
Bacteria	Thermotogales	Thermotoga	<i>Thermoplasma volcanium</i>
Bacteria	Eubacteria	Spirochaetales	<i>Thermotoga maritima</i>
Bacteria	Firmicutes	Grupo Bacillus/ Clostridium	<i>Treponema pallidum</i>
Bacteria	Proteobacteria	Subdivisión gama	<i>Ureaplasma urealyticum</i>
Bacteria	Proteobacteria	Subdivisión gama	<i>Vibrio cholerae</i>
Bacteria	Proteobacteria	Subdivisión gama	<i>Xylella fastidiosa</i> 9a5c
Bacteria	Proteobacteria	Subdivisión gama	<i>Yersinia pestis</i>

BIBLIOGRAFÍA

- 1.- **Ulrich Dobrindt, Jorg Hacker.** Whole genome plasticity in pathogenic bacteria. *Current Opinion in Microbiology.* 2001; 4: 550-557.
- 2.- **Brendan W. Wren.** Microbial genome analysis: insights into virulence, host adaptation and evolution. *Nature.* 2000; 1: 30-39.
- 3.- **Lawrence J.G.** Gene transfer, speciation, and the evolution of bacterial genomes. *Current Opinion in Microbiology.* 1999; 2: 519-523.
- 4.- **Aravind L., Tatusov R. L., Wolf Y. I., Walker D. R., Koonin E. V.** Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet.* 1998; 14: 442-444.
- 5.- **Lawrence J.G.** Selfish operons and speciation by gene transfer. *Trends Microbiol.* 1997; 5: 355-359.
- 6.- **Ochman, H., Lawrence J. G., & Groisman, E. A.** Lateral gene transfer and the nature of bacterial innovation. *Nature.* 2000; 405:299-304.
- 7.- **Jain R., Rivera M. C., Lake J. A.** Horizontal gene transfer among genomes: complexity hypothesis. *Proc Natl Acad Sci USA.* 1999; 96: 1971-1976.
- 8.- **Lawrence J. G., Ochman H.** Reconciling the many faces of lateral transfer. *Trends in Microbiol.* 2002; 1: 1-39.
- 9.- **Eisen J. A.** Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Current Opinion in Genetics & Development.* 2000; 10: 606-601.
- 10.- **Koonin E. V., Kira S. Makarova, and L. Aravind.** Horizontal gene transfer in Prokaryotes: Quantification and Classification. *Annu. Rev. Microbiol.* 2001; 55: 709-742.
- 11.- **Eisen J. A.** Assessing evolutionary relationships among microbes from whole-genome analysis. *Current Opinion in Microbiology.* 2000; 3:475-480.

- 12.- **Fitch W. M.** Homology, a personal view on some of the problems. *TIG.* 2000; 16: 227-231.
- 13.- **Olsen G. J. and Woese C. R.** Archaeal Genomics: An Overview. *Cell.* 1997; 89: 991-994.
- 14.- **Yang Z. and Bielawski J. P.** Statistical methods for detecting molecular adaptation. *TREE.* 2000; 15: 496-503.
- 15.- **Yang Z.** PALM: a program packega for phylogenetic analysis by maximum likelihood. *CABIOS.* 1997; 13:555-556(<http://abacus.gene.ucl.ac.uk/software/paml.html>).
- 16.- **Lawrence J. G. and Ochman H.** Molecular archaeology of the Escherichia coli genome. *Proc. Natl. Acad. Sci.* 1998; 95: 9413-9417.
- 17.- **Joyce E. A., Chan K., Salama N. R. and Flakow S.** Redefining bacterial populations: a post-genomic reformation. *Nature.* 2002; 3: 462-473.
- 18.- **Yang Z., Swanson W. J. and Vacquier V. D.** Maximum-Likelihood Analysis of Molecular Adaptation in Abalone Sperm Lysin Reveals Variable Selective Presures Among Lineages and Sites. 2000; 17: 1446-1455.
- 19.- **Gibas C. & Jambeck P.** 2001. *Developing Bioinformatics Computer Skills.*(ed.) O'REILLY, Sebastopol.
- 20.- **Huelsenbeck J. P. and Rannala B.** Phylogenetic Methods Come of Age: Testing Hypotheses in an Evolutionary Context. *Science.* 1997; 276: 227-232.
- 21.- **Jordan I. K., Rogozin I. B., Wolf Y. I. and Koonin E. V.** Essential Genes Are More Evolutionarily Conserved Than Are Nonessential Genes in Bacteria. *Genome Research.* 2002; 12: 962-968.
- 22.- **Hirsh A. E. & Fraser H. B.** Protein dispensability and rate of evolution. *Nature.* 2001; 411: 1046-1049.

GLOSARIO

Clado: Linaje filogenético de taxas emparentados cuyo origen es un ancestro común.

Descendencia vertical: Transferencia de genes de progenitor a descendiente.

Descendencia horizontal: Transferencia horizontal.

Especies químeras: Especies cuyo material genético proviene de más de una especie.

Filogenia: Relación histórica que hay entre los linajes de los organismos.

Genes informacionales: Genes implicados en los sistemas de procesamiento de información -replicación, transcripción y traducción-.

Genes operacionales: Genes implicados en el metabolismo.

Homología: Común ancestría de dos o más genes, o productos de los genes.

Ortología: Homología que se origina por especiación.

Paralogía: Homología que se origina por duplicación de genes.

Selección negativa ó purificadora: Selección natural que actúa contra las mutaciones deletéreas con coeficientes de selección negativos.

Selección positiva ó diversificadora: Selección que fija mutaciones ventajosas con coeficientes de selección positivos.

Sesgo en el uso de codones: frecuencias desiguales en los codones de un gen; uso preferencial de ciertos codones en lugar de otros.

Sesgo en el ratio de transición/transverción: tasas de sustitución desiguales entre los nucleótidos, la más alta de las cuales es para las transiciones.

Sustitución sinónima: Sustitución de un nucleótido que no resulta en un cambio de aminoácido.

Sustitución no sinónima: Sustitución de un nucleótido que resulta en un cambio de aminoácido.