

00377



UNIVERSIDAD NACIONAL AUTONOMA
DE MEXICO

POSGRADO EN CIENCIAS
BIOLOGICAS

Facultad de Ciencias

Estudio de las Principales Mutaciones Involucradas en la
Evolución de las Rutas Metabólicas

T E S I S

QUE PARA OBTENER EL GRADO ACADEMICO DE
MAESTRO EN CIENCIAS BIOLOGICAS

(Biología Experimental)

P R E S E N T A

Biol. Diego Claudio Cortez Quezada

Director de Tesis: Dr. Antonio Eusebio Lazcano-Araujo Reyes

México, D. F.



Octubre, 2004

COORDINACIÓN

TESIS CON
FALLA DE ORIGEN



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

ESTA TESIS NO DEBE
SALIR DE LA BIBLIOTECA

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
INSTITUTO DE INVESTIGACIONES Y ENSEÑANZA EN QUÍMICA
CARRERA DE QUÍMICA
MÉXICO, D.F. 1980



UNIVERSIDAD NACIONAL
AUTÓNOMA DE
MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS COORDINACIÓN

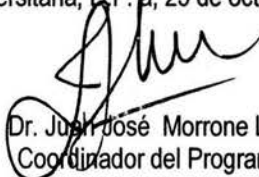
Ing. Leopoldo Silva Gutiérrez
Director General de Administración Escolar, UNAM
Presente

Por medio de la presente me permito informar a usted que en la reunión ordinaria del Comité Académico del Posgrado en Ciencias Biológicas, celebrada el día 11 de octubre del 2004, se acordó poner a su consideración el siguiente jurado para el examen de grado de Maestría en Ciencias Biológicas (Biología Experimental) del(a) alumno(a) **Cortez Quezada Diego Claudio**, con número de cuenta 98527832 con la tesis titulada: **"Estudio de las principales mutaciones involucradas en la evolución de las rutas metabólicas"**, bajo la dirección del(a) **Dr. Antonio Eusebio Lazcano-Araujo Reyes**.

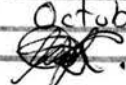
Presidente:	Dr. Diego González Halphen
Vocal:	Dr. Enrique Merino Pérez
Secretario:	Dr. Antonio Eusebio Lazcano-Araujo Reyes
Suplente:	Dr. Lorenzo Patrick Segovia Forcella
Suplente:	Dr. Arturo Carlos Il Becerra Bracho

Sin otro particular, quedo de usted.

Atentamente
"POR MI RAZA HABLARA EL ESPIRITU"
Cd. Universitaria, D.F. a, 25 de octubre del 2004


Dr. Juan José Morrone Lupi
Coordinador del Programa

c.c.p. Expediente del interesado

Autorizo a la Dirección General de Bibliotecas de la UNAM a difundir en formato electrónico e impreso el contenido de mi trabajo recepcional.
NOMBRE: Diego Claudio Cortez Quezada
FECHA: 29 Octubre 2004
FIRMA: 

Agradezco a **CONACYT**, beca número 172367, por el apoyo otorgado para la realización de esta tesis.

Agradezco enormemente a los miembros de mi comité tutorial, **Dr. Diego González Halphen, Dr. Lorenzo Segovia Forcella y Dr. Antonio Lazcano-Araujo Reyes** por sus valiosas aportaciones a este trabajo.

A mi madre. Por la inimaginable falta que me haces.

**A mi hermanito, mi papá y Paulita.
Porque sin ustedes desaparezco.**

A los cuates que ahí siguen.

"There is grandeur in this view of life, with its several powers, having been originally breathed by the Creator into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being evolved."

Charles Darwin, last paragraph of origin of species (1859).

*"Ya había soñado todos mis grandes sueños.
No me habían llevado a ninguna parte,
y ahora estaba demasiado cansado
para concebir unos nuevos".*

Paul Auster.

INDICE

Resumen	1
Abstract	2
Introducción	3
La Importancia de Llamarse Genómica	3
La Evolución de las Rutas Metabólicas y la Duplicación de Genes	3
Rearreglos Genómicos: Porque No sólo las Duplicaciones son Importantes	7
Las Bacterias como Modelo de Estudio	8
Métodos	10
Bases de Datos	10
Secuencias Simples	10
Transferencia Horizontal	12
Duplicación de Fragmentos de DNA	14
Pérdida de Material Genético	15
Genes sin Pasado Evidente y Pérdida en la Capacidad de Detección de GTH	15
Resultados y Discusión	16
Secuencias Simples (SeSp)	16
Transferencia Horizontal	22
Detectando Genes Introducidos Artificialmente	23
Aplicación del Modelo de Markov para Detectar los Genes Recientemente Transferidos en el Genoma de <i>E. coli</i> K12 MG1655, <i>E. coli</i> O157 EDL933 y <i>S. typhimurium</i> LT2	27
Duplicación de Fragmentos de DNA	30
Pérdida de Material Genético	35
Desentrañando la Historia de los Genes sin Pasado Evidente y como Aprendí a Extraviar la Información	38

Conclusiones	41
Sobre las Secuencias Simples	41
Sobre la Transferencia Horizontal	42
Sobre las Duplicaciones de Genes	43
Sobre la Pérdida de Material Genético y la Reconstrucción de Historias	44
Sobre la Evolución de las Rutas Metabólicas y los Procariontes que las Poseen	45
Referencias	47
Apéndice 1. Tabla de Programas Más Importantes Escritos en PERL	50

RESUMEN

La evolución de las rutas metabólicas es un tema que ha captado el interés de los investigadores durante décadas. El presente estudio plantea una aproximación al problema a partir de la valoración de diferentes tipos de mutaciones, con el fin de identificar los sesgos que han existido en la selección de dichas mutaciones en las rutas metabólicas. Las mutaciones analizadas fueron: aparición de secuencias simples, duplicación de fragmentos de DNA, pérdida de material genético y transferencia horizontal. La investigación se realizó en tres genomas modelo: las enterobacterias *Escherichia coli* cepas K-12 MG1655 y O157 EDL933 y *Salmonella typhimurium*. Se elaboraron algoritmos computacionales que permitieron la correcta detección de las mutaciones. En particular, se construyó el modelo más eficiente de detección de genes transferidos reportado hasta la fecha. La eficiencia de los modelos de detección de transferencia horizontal fue evaluada a través de un experimento control innovador que consiste en la adición artificial de genes foráneos a un genoma. Los resultados del trabajo muestran que las secuencias simples están presentes principalmente en proteínas de membrana y, en algunos casos, su aparición se remonta muy atrás en el tiempo lo que sugiere que dicho fenómeno estaba presente en etapas tempranas de la vida. La duplicación de genes es un fenómeno que se ha presentado en todas las rutas metabólicas, a lo largo de toda la historia evolutiva de los organismos. Hay rutas metabólicas donde las duplicaciones datan de épocas muy remotas y rutas donde las duplicaciones no han dejado de aparecer. Las rutas auxiliares (metabolismo de xenobióticos, transporte a través de membranas) son las que se han mantenido en constante cambio, aceptando duplicaciones, transferencias horizontales y pérdidas de genes. Las copias recientes de genes son la materia prima en donde se pueden ensayar nuevas funciones. Aquellos genes que signifiquen una ventaja para el organismo en determinado ambiente es muy probable que sean difundidos lateralmente entre las poblaciones bacterianas. La convergencia de una nueva vía favorable en una sola población de bacterias determina que dicha vía no esté sujeta a la transferencia horizontal; se vuelve inmutable. Hay funciones celulares, sin embargo, que nunca se estabilizan: aquellas cuyas proteínas median la interacción del organismo con su medio ambiente (proteínas de membrana, proteínas de respuesta a estrés, enzimas de las rutas de xenobióticos, etc.). Son los genes que codifican para este tipo de proteínas los que permiten que las bacterias se adapten al ambiente.

ABSTRACT

The evolution of metabolic pathways has been an important issue during the last decades. This thesis presents an approximation to the problem through the evaluation of different types of mutations, aiming to clarify the biases that have existed in the detection of these mutations in the metabolic pathways. The mutations analyzed were: simple sequences appearance, DNA fragments duplications, loss of genetic material, and horizontal transfer. The research was made in three genomes: *Escherichia coli* strains *K-12 MG1655* and *O157 EDL933* and *Salmonella typhimurium*. Computational algorithms were developed in order to detect in an accurate manner, the mutations. Particularly, I constructed the most efficient model for horizontal transfer gene detections, made to the date. The efficiency of the horizontal transfer detection models was evaluated through an experimental control based in the artificial additions of foreign genes into a given genome. The results show that the simple sequences are present mainly in membrane proteins, and in some cases, its origin could be ancient, meaning this phenomenon was present in the early stages of life. The gene duplication is a phenomenon that has appeared in all the metabolic pathways, along the evolutionary history of the organisms. There are some metabolic pathways where the duplications are ancient and some pathways that have kept accepting duplications. The auxiliary pathways (such as xenobiotics metabolism and membrane transport) have maintained constant change, accepting duplications, horizontal transfers and gene loss. The recent copies of genes are the raw material where new functions can be tested. Those genes that represent an advantage for the organism in certain environments are extremely probable to be laterally spread to different bacterial populations. The convergence of some favorable pathway in one bacterial population determines that such pathway will not be subject to horizontal transfer; it becomes fixed. However, there are some cellular functions that never get stabilized: those involved in the interaction between the organism and its environment (such as xenobiotics metabolism, shock proteins and membrane transport). The genes that code for these kinds of proteins are those that allow the bacteria to get adapted to the environment.

INTRODUCCIÓN

“¿Por qué no ocurren hoy, se pregunta, esos milagros de que se habla como sucesos pasados? Yo podría contestar que eran necesarios, antes de que el mundo creyera, para llevarle a creer; pero cualquiera que busque hoy prodigios para despertar su fe es, por su parte, un gran prodigio al negarse a creer lo que todo el mundo cree”.

San Agustín.

La Importancia de Llamarse Genómica

“Las religiones, como las luciérnagas, necesitan de la oscuridad para brillar”.

Schopenhauer.

En el genoma se encuentran albergados todos los genes que codifican para un organismo. Su estudio es por ende, un acercamiento a la historia evolutiva de ese ser vivo. Entender cómo en un organismo se han acumulado y seleccionado variaciones de acuerdo a las condiciones ambientales en donde se ha desarrollado, es una tarea con implicaciones invaluable, particularmente en lo referente a la evolución de rutas metabólicas; Este ha sido el objetivo principal de este trabajo.

La Evolución de las Rutas Metabólicas y la Duplicación de Genes

“Muchos hombres que hoy están dispuestos a dejarse matar por defender un milagro, lo hubieran puesto en duda si hubiesen estado presentes al producirse”.

Lichtenberg.

En 1936 Oparin dio a conocer sus ideas sobre el origen heterotrófico de la vida. En ellas planteaba una síntesis prebiótica de compuestos orgánicos en una atmósfera reductora. La sucesiva interacción de estos compuestos a lo largo del tiempo y la acumulación de moléculas, producto de las reacciones espontáneas, conformarían las primeras rutas metabólicas.

Tras las propuestas de Oparin, se requirieron muchos años y otros tantos descubrimientos científicos para que fuera presentada la primera hipótesis formal de evolución de las rutas metabólicas. Este logro intelectual fue hecho por Horowitz, quien en 1945 presentó la conocida hipótesis retrógrada (Horowitz 1945). En ella, una primera enzima cataliza cierta reacción. Una duplicación de dicha enzima y su subespecialización, permiten la catálisis de una nueva reacción cuyo producto es precisamente el sustrato de la primera enzima. De esta forma, y tras sucesivas duplicaciones, se construye una vía metabólica cuyo inicio es precisamente el último producto de la vía (Figura 1). Muchos de los fundamentos de la hipótesis de Horowitz no han podido ser demostrados; por el contrario, se ha encontrado que la inmensa mayoría de las enzimas que están en una misma ruta metabólica catalizan reacciones muy distintas

y, además, no son homólogas (Lazcano y Miller 1999).

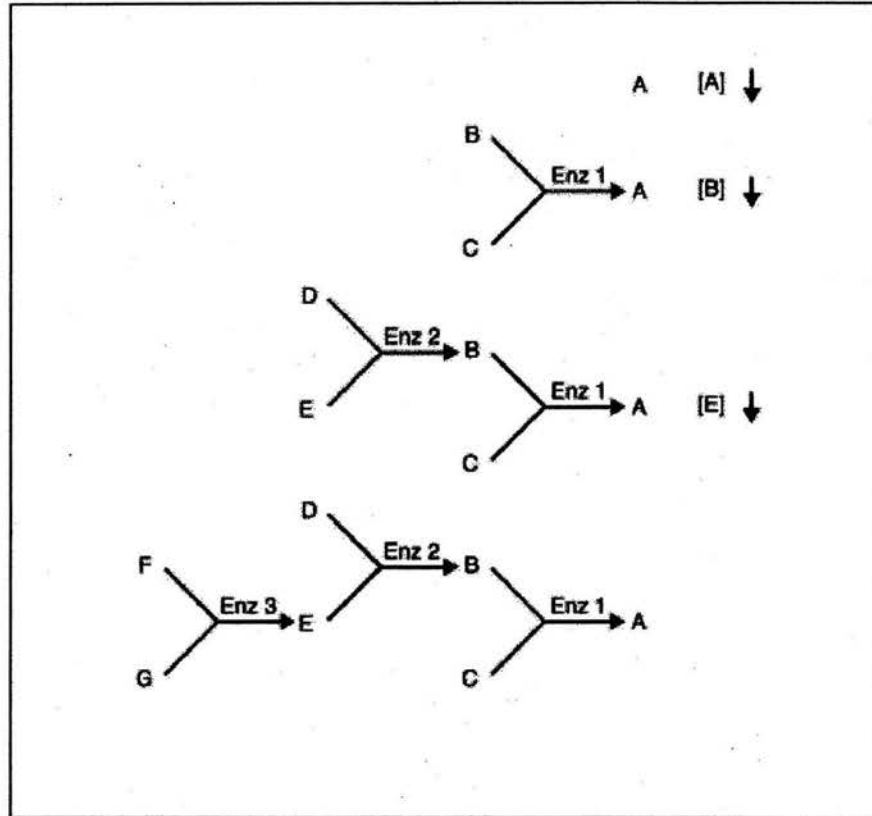


Figura 1. Hipótesis Retrógrada (figura tomada de Rison y Thornton 2002).

Debido a todos los problemas teóricos y prácticos que se encontraron en la tesis de Horowitz, en 1976 Jensen (Jensen 1976) y varios más de manera independiente (Ycas, M 1974) postularon la hipótesis de Patchwork. En ella, se planteaba que una enzima poco específica y, por lo tanto, poco eficiente, catalizaba varias reacciones muy similares. Sucesivas duplicaciones y subespecializaciones hacia una sola de las múltiples funciones ancestrales, dieron como resultado una serie de proteínas homólogas, cada una de las cuales catalizaría una única reacción que llevaba a cabo la enzima ancestral. Suponiendo que este fenómeno ocurriera en diversas enzimas, el resultado final sería la aparición de vías metabólicas con enzimas especializadas, homólogas entre sí, pero que no necesariamente se encontrarían en la misma vía o en pasos consecutivos de la misma (Figura 2). El principal problema con esta hipótesis es que para que sea operable es necesaria la existencia de una biosíntesis de proteínas bien establecida, es decir, sólo pudo ocurrir después de la aparición del mundo de RNA y proteínas (Lazcano y Miller 1999).

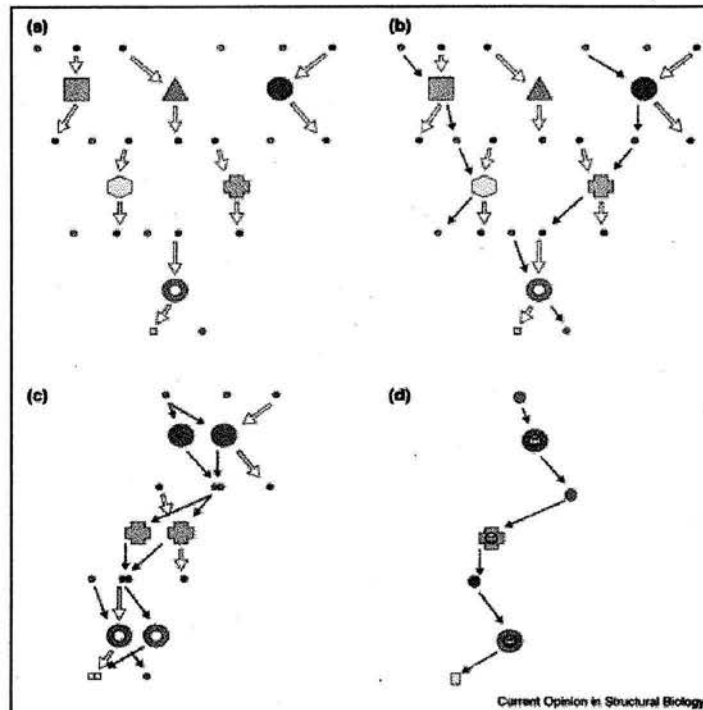


Figura 2. Hipótesis de Patchwork (figura tomada de Rison y Thornton 2002).

La necesidad de explicar el inicio de las primeras rutas metabólicas, previo a la biosíntesis de proteínas, condujo a que en 1999 Lazcano y Miller (Lazcano y Miller 1999) propusieran la hipótesis del origen semienzimático de las rutas metabólicas. Ellos asumieron que: en un inicio existieron compuestos prebióticos estables disponibles en los océanos primitivos, así como compuestos resultados de las rutas rudimentarias catalizadas por ribozimas. La duplicación de genes proporcionó enzimas no específicas. Finalmente, las primeras reacciones no-enzimáticas reclutaron dichas enzimas.

Las propuestas evolutivas del origen de las rutas metabólicas proponen a la duplicación de genes como principal fenómeno evolutivo involucrado en la estructuración de las rutas metabólicas. Este hecho cobra fuerza a partir de los estudios de los genomas totalmente secuenciados. Los resultados de los análisis muestran que un ~40% de los genes dentro de un genoma tiene por lo menos un gen parálogo (Hooper y Berg 2003a). ¿Es entonces la duplicación de fragmentos de DNA la manera más común de generar variabilidad en un genoma? Según Bernard y Riley (1995), sí y explican que una duplicación de material genético genera nuevas copias de genes cuya función ya ha sido probada y previamente seleccionada al interior de ese genoma y lo único que quedaría por seleccionarse es la capacidad del organismo de soportar la amplificación en esa actividad.

Por muchos años se pensó que el modelo de evolución de genes duplicados propuesto por Ohno en 1970 (Ohno 1970) era el correcto. En éste, una de las copias de algún gen recientemente duplicado no se encuentra bajo ninguna presión de selección y es libre de acumular mutaciones al azar. De esta manera, diverge hasta obtener una nueva función. Cuando la adquiere, vuelve a estar sujeto a un proceso de selección.

En realidad, existen tres destinos para los genes duplicados: selección, neutralidad o pérdida. En la teoría original de Ohno, la neutralidad es obligatoria para los genes duplicados independientemente de su destino. Recientemente varios autores han cuestionado este periodo de neutralidad génica (Kondrashov et al. 2002; Hooper y Berg 2003a). Estos trabajos muestran que tal condición puede ser mínima o inexistente. Los genes que se duplican tienen sólo dos destinos posibles: en el primero, se seleccionan de manera que se amplifique la función que presentan; en el segundo escenario son rápidamente contra seleccionados. ¿De qué depende que se amplifique una cierta función o se contra seleccione? Se ha propuesto que los genes con funciones débiles o auxiliares son los que más se duplican porque presentan una presión de selección menos intensa que aquellos genes de funciones fuertes o bien establecidas (Kondrashov et al. 2002; Hooper y Berg 2002; Hooper y Berg 2003a). Un ejemplo que se ha señalado a este respecto es la aparentemente acelerada tasa de duplicaciones que tienen los genes de transferencia horizontal, cuya presión de selección podría ser mínima en algunos casos. Algunas investigaciones parecen corroborar este supuesto (Hooper y Berg 2003b) ya que se ha encontrado que algunos genes claramente transferidos horizontalmente presentan algunos parálogos dentro del genoma hospedero.

Los mecanismos mejor conocidos de generación de duplicaciones en un genoma son: entrecruzamiento desigual (CD), escisión y reinserción del círculo (ERC), y el círculo rodante (CR) (Romero y Palacios 1997; Figura 3). Estos mecanismos producen de igual manera secuencias simples, secuencias cortas en *tandem*, duplicación de fragmentos grandes de DNA o la pérdida de material genético.

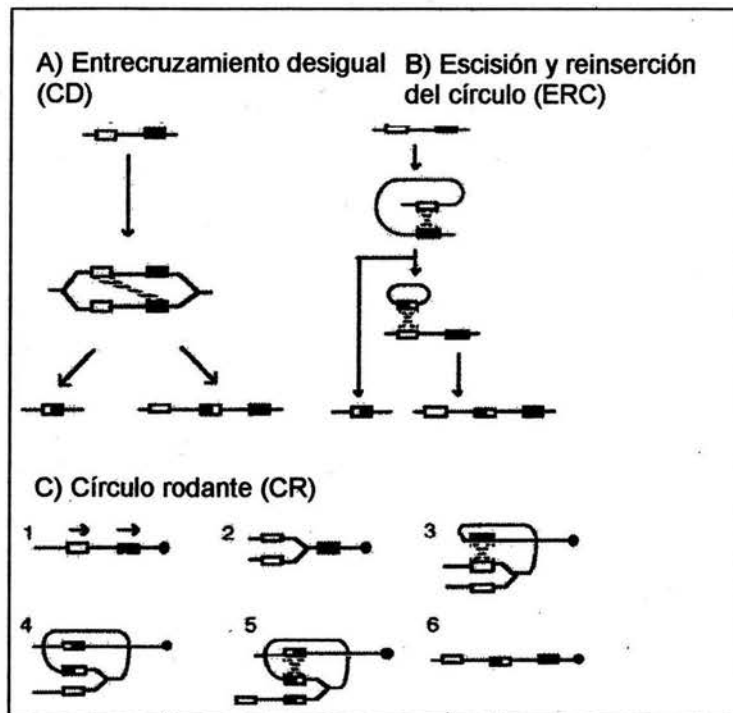


Figura 3. Mecanismos de generación de duplicación en procariontes (figura tomada de Romero y Palacios 1997).

Rearreglos Genómicos: Porque No sólo las Duplicaciones son Importantes.

"Yo amo a la sabiduría más de lo que ella me ama a mí".

Lord Byron.

Aunque se ha postulado que la duplicación de fragmentos de material genético podría haber sido el fenómeno evolutivo más importante en la estructuración de las rutas metabólicas y de los genomas, existen otro tipo de rearrreglos genómicos de gran importancia que, por ser más difíciles de evaluar, su estudio ha sido estudiado en menor grado. Aún así, es indispensable conocerlas y cuantificarlas para conocer de una manera correcta la evolución de los genomas. Estos rearrreglos incluyen a la aparición de secuencias simples, la transferencia horizontal y la pérdida de material genético.

Las secuencias simples. Son regiones del genoma que presentan sesgos en la composición hacia algún aminoácido o nucleótido en particular (Wootton y Federhen 1993). Se han propuesto múltiples mecanismos para explicar la aparición de este tipo de secuencias: mal apareamiento y deslizamiento de la hebra (Bzymek & Lovett 2001), entrecruzamiento desigual (por recombinación de homólogos) e inserción de círculo rodante (Romero y Palacib 1997). Se han hecho varios trabajos analizando su función y selección. Se les ha relacionado con el origen y diversificación de la patogenicidad en bacterias (Baylis, Dixon y Moxon 2004), han sido postuladas como mediadoras de la regulación genética (Gur-Arie *et al.* 2000) y también como generadoras de material genético *de novo* dentro de los genomas (Becerra, Cocho, Delaye y Lazcano, en prep).

La transferencia horizontal. La transferencia horizontal de genes se ha supuesto como una de las principales fuerzas evolutivas que moldean los genomas en los procariontes. La dificultad que se desprende de su análisis ha generado una gran controversia a lo largo de los últimos años. Se han realizado innumerables estudios que han intentado revelar la magnitud del fenómeno y la naturaleza de los genes que están involucrados (Aravind *et al.* 1998; Hayes & Borodovsky 1998; Kyrpides *et al.* 1999; Ochman & Lawrence 1998; García-Vallvé *et al.* 2000; Koonin 2001; Hooper & Berg 2002; Snel *et al.* 2002; Daubin, Moran & Ochman 2003). La construcción de un modelo de detección que pueda distinguir entre pérdida de genes y transferencia horizontal sería una herramienta analítica sumamente poderosa, necesaria e importante.

Se han utilizado dos enfoques distintos para detectar los genes transferidos horizontalmente. El primero de ellos se basa en el análisis de los usos de codones y los porcentajes de GC presentes en un cierto genoma. De esta forma se determina qué genes presentan una composición anormal y por lo tanto no pertenecen al genoma (Ochman y Lawrence 1998; Garcia-Vallvé *et al.* 2000). La segunda idea utiliza comparaciones de contenidos génicos entre organismos cercanos filogenéticamente. Se construye un perfil de distribuciones (PD) de presencia-ausencia en estos organismos para cada gen (Ragan y Charlebois 2002; Daubin, Lerat y Perrière 2003; Daubin, Moran y Ochman 2003). Cuando un gen no se encuentra en ninguno de los genomas cercanos filogenéticamente, usando varios umbrales estrictos, se postula como un gen que probablemente fue transferido horizontalmente.

Recientemente, los modelos composicionales han sufrido problemas de credibilidad porque se ha comprobado que no hay diferencias significativas en la composición de bases y el uso de codones entre los genes propuestos como transferidos horizontalmente y los genes conservados (Koski et al. 2001; Wang 2001). Estos resultados podrían deberse a la marcada heterogeneidad composicional presente en cualquier genoma (Guindon y Perrière 2001; Daubin y Perrière 2003). Aunado a estos resultados, Ragan (2001) encontró que, al ser aplicados cuatro diferentes métodos de detección de genes de transferencia horizontal en el genoma de *Escherichia coli* K12, cada uno de ellos detectaba un grupo muy distinto de genes. Ragan concluyó en su trabajo que los resultados se debían a que los métodos composicionales estaban detectando genes transferidos recientemente, mientras que los otros métodos detectaban genes más antiguos.

Por otro lado, hemos sido testigos de graves errores científicos en el área. Por ejemplo, el cometido por los miembros del proyecto del genoma humano, quienes declararon que habían detectado 113 genes de bacterias que habían llegado por transferencia horizontal al linaje de los vertebrados (Salzberg et al. 2001). Esta aseveración fue rápidamente refutada por numerosos trabajos (Stanhope et al. 2001; Roelofs y Van Haastert 2001). Lo mismo ocurrió con la propuesta de Aravind *et al* (1998) de un intenso intercambio de genes entre bacterias y arqueas hipertermofílicas. Nuevamente, las propuestas de este trabajo fueron impugnadas con éxito (Krypides 1999).

Aunque se han teorizado gran número de modelos de detección de genes transferidos horizontalmente, es factible considerar una pregunta que hasta el momento no ha sido planteada: ¿estas metodologías comúnmente utilizadas están realmente detectando genes de transferencia horizontal? La respuesta no es clara. Aún no hemos sido capaces de construir modelos y controles experimentales que nos aseguren la detección de aquellos genes que han tenido un origen por transferencia horizontal.

La pérdida de material genético. Los mismos mecanismos que generan la aparición de secuencias duplicadas, pueden, en ciertos casos, favorecer la pérdida de fragmentos de DNA (Romero y Palacios 1997). Algunos organismos, especialmente parásitos, han presentado un proceso de pérdida de genes muy acentuado (Moran 2003). Aunque se ha planteado que en ellos la pérdida de genes se ha producido por la erosión gradual de *loci* y operones individuales.

Las Bacterias como Modelo de Estudio

"Una cosa es desear que la verdad esté de nuestro lado, y otra muy distinta desear sinceramente estar del lado de la verdad".
Ricard Whately.

Los organismos más sencillos morfológicamente que habitan este mundo son los procariontes. Por lo tanto, el análisis de los procesos y cambios que les han ocurrido implican también metodologías más sencillas. El estudio de estos organismos permitirá entender de manera más cercana a la realidad, la dinámica interna actual de sus genomas y generará también un bosquejo de su historia evolutiva.

Este trabajo plantea el análisis global de los genomas completos de las enterobacterias con el fin de

distinguir todas las mutaciones que han ocurrido y se han seleccionado a lo largo del tiempo. Se sustentará en la acumulación diferencial de las mutaciones en las rutas metabólicas, con el fin de vislumbrar su historia evolutiva.

Los modelos de estudio de este trabajo fueron las enterobacterias: *Escherichia coli* cepas K-12 MG1655 y O157 EDL933 y *Salmonella typhimurium*. Se escogieron como objeto de estudio debido a la vasta cantidad de información metabólica que se tiene de ellas: la gran mayoría de sus genes tienen una asignación funcional y espacial dentro de las rutas metabólicas, información indispensable para el tipo de investigación que se plantea.

MÉTODOS

*"El pensamiento llega cuando él quiere,
no cuando quiere uno".
Nietzsche.*

Bases de Datos. Dos diferentes bases de datos fueron requeridas. La primera de ellas estaba compuesta por los genomas en nucleótidos y aminoácidos de las siguientes bacterias en formato Fasta: *E.coli K-12 MG1655*, *E.coli K-12 W3110*, *E.coli O157 EDL933*, *E.coli O157 Sakai*, *E.coli CFT073*, *S.typhi CT18*, *S.typhimurium*, *Y.pestis CO92*, *Haemophilus influenzae*, *Neisseria meningitidis MC58 (serogroup B)*, *Campylobacter jejuni*, *Mesorhizobium loti*, *Caulobacter crescentus*, *Bacillus subtilis*, *Bacillus halodurans*, *Staphylococcus aureus N315 (MRSA)*, *Clostridium acetobutylicum*, *Mycoplasma genitalium*, *Mycobacterium tuberculosis H37Rv (lab strain)*, *Chlamydomydia pneumoniae CWL029*, *Deinococcus radiodurans*, *Aquifex aeolicus*, *Thermotoga maritima*, *Methanococcus jannaschii*, *Thermoplasma volcanium*, *Sulfolobus solfataricus* y *Aeropyrum pernix*. Todos los genomas se obtuvieron del servidor del KEGG (<http://www.genome.ad.jp/kegg/kegg2.html>).

La segunda base de datos contenía la información metabólica y la clasificación de todos los genes de los genomas de *E.coli K-12 MG1655*, *E.coli O157 EDL933* y *S.typhimurium*.

Se hicieron comparaciones de los contenidos génicos entre *E.coli* cepas *K-12 MG1655*, *K-12 W3110*, *O157 EDL933*, *O157 Sakai*, *CFT073*, las especies *S.typhi CT18* y *S.typhimurium* y *Y.pestis CO92*, utilizando el programa Blastp (Altschul et al. 1997) con un límite superior para el valor de e de 0.001. Las mejores parejas bidireccionales entre dos genomas fueron tomadas como genes ortólogos.

Secuencias Simples. El programa SEG fue utilizado para buscar secuencias de baja complejidad en aminoácidos. Los valores de entrada fueron: A) análisis por ventanas de 12 aminoácidos. B) corte bajo de 1.9, este es el parámetro inicial K(1) en la ecuación de Wootton y Federhen (Wootton y Federhen 1993; Wootton 1994). En la primera etapa del algoritmo, SEG busca ventanas con un valor igual o menor a este dato, entre menor sea el valor de este parámetro más simple deberá ser la secuencia. C) corte alto o parámetro de complejidad de la extensión del segmento de 2.5; en la segunda parte del algoritmo, SEG extiende el segmento que encontró, permitiendo que presente una mayor complejidad para detectar secuencias simples crípticas o más deterioradas; los valores parten del valor designado en el corte bajo.

El programa SEG se aplicó a *E.coli K-12 MG1655*, *E.coli O157 EDL933* y *S.typhimurium*.

Los genes que presentaron secuencias simples se separaron en una base de datos diferente. Las secuencias simples se contabilizaron, se calculó su frecuencia de aminoácidos y la cobertura en el gen. Se buscó la asignación metabólica a todos los genes que presentaron SeSp. Se calculó la frecuencia de aminoácidos por ruta metabólica, y finalmente, se estudió caso por caso las anotaciones de cada uno de los genes con SeSp.

La ancestría de los genes con secuencias simples se analizó buscando su presencia en los genomas de las bacterias: *E.coli* cepas *K-12 MG1655*, *K-12 W3110*, *O157 EDL933*, *O157 Sakai*, *CFT073* y las especies *S.typhi CT18*, *S.typhimurium LT2* y *Y.pestis CO92*. Los genes con secuencias simples que

estuvieron sólo en un genoma, se depuraron eliminando todos aquellos genes que tuvieran un parólogo conservado o en aquellos casos cuyos ortólogos se hubieran perdido en los otros linajes. A los genes resultantes se les consideró genes con SeSp seleccionadas *de novo* recientemente.

Se escogieron 50 genes al azar que estuvieran compartidas entre *E.coli K-12 MG1655*, *E.coli O157 EDL933* y *S.typhimurium LT2*. Se alinearon las tres secuencias y se analizaron los cambios en las firmas de las secuencias simples.

Con el programa Statistica 7.0 se realizó una X^2 de esperados contra observados de la distribución de las SeSp en las diferentes rutas metabólicas. Los esperados se calcularon: [#total de SeSp en el genoma] * [#total de genes en cada vía metabólica] / [#total de genes en el genoma]. Los valores significativos tuvieron que ser $p < 0.001$. También se realizó una X^2 del número de genes relacionados a la patogenicidad que deberían presentar secuencias simples para que este dato fuera significativo. Los datos esperado se calcularon: [#total de SeSp en el genoma] * [#total de genes relacionados con la patogenicidad] / [#total de genes en el genoma]. Los valores significativos tuvieron que ser $p < 0.001$. Finalmente, se estudiaron las distribuciones funcionales de los genes con SeSp entre las dos cepas de *E.coli*. Se hizo una nueva X^2 de los valores observados en *E.coli K12 MG1655* y los valores esperados para la misma bacteria de acuerdo a los datos obtenidos de *E.coli O157 ED933*. Los valores esperados se calcularon: [#total de SeSp en el genoma de *E.coli O157*] * [#total de genes en cada vía metabólica de *E.coli K12*] / [#total de genes en el genoma de *E.coli O157*]. Los valores significativos tuvieron que ser $p < 0.001$.

Las proteínas de *E.coli K12* con estructura cristalográfica que, además, presentan SeSp son: Proteínas ribosomales L20, L18, L23, L1 y L7/L12; Factores de iniciación de la traducción IF-2 e IF-3; citocromo oxidasa; citocromo b561; subunidad III de la ubiquinol oxidasa; tRNA sintetasas de valina, cisteína y la subunidad β de la glicina; factores σ 70 y 54; RNA polimerasa subunidad Ω ; polimerasa III subunidades δ , δ' y τ - γ ; DNA polimerasa I; ATP sintetasas proteína I; ATP sintetasas proteína C. Las estructuras se localizaron en la página del PDB (<http://www.rcsb.org/pdb/>). Se localizaron a través del KEGG todos los ortólogos de cada una de estas proteínas. Se alinearon usando el programa tcoffee (<http://igs-server.cnrs-mrs.fr/Tcoffee/tcoffee.cgi/index.cgi>) y se visualizaron con Pfaat (<http://pfaat.sourceforge.net/>). Se analizó, a partir de la secuencia simple de *E.coli K12*, el mantenimiento de las firmas en todos los linajes. La conservación del hecho de presentar SeSp se hizo también a partir de la alineación. Se ubicó la región correspondiente a la SeSp de *E.coli K12* y se hizo una búsqueda con el programa SEG, cuyos parámetros fueron 12-1.9- 2.5, de la presencia de una secuencia simple.

Los análisis de predicción de estructura secundaria se llevaron a cabo con el programa PHDsec a través de su página de internet (<http://maple.bioc.columbia.edu/predictprotein/>; Rost y Sander 2000). Debido a que este programa sólo trabaja por red, es muy lento y se compite con otros usuarios para que se haga el análisis, se tuvo que restringir a 25 el número de proteínas analizadas de cada grupo: Proteínas de membrana, enzimas y proteínas no-asignadas a ninguna categoría. Las proteínas fueron seleccionadas aleatoriamente a través de un algoritmo escrito en perl. La predicción de los dominios intermembranales se hizo con el programa TMHMM a través del servidor (<http://www.cbs.dtu.dk/services/TMHMM-2.0>). Se

contabilizó para cada grupo el número de veces que una secuencia simple estaba dentro del dominio intermembranal.

Transferencia Horizontal. Para probar los distintos métodos de detección de genes transferidos horizontalmente, se introdujeron en genoma de *E.coli K12 MG1655* y *E.coli O157 EDL933* cien genes obtenidos aleatoriamente de los siguientes organismos con genoma totalmente secuenciado obtenidos de la base de datos del KEGG (<http://www.genome.ad.jp/kegg/kegg2.html>): *Y.pestis CO92*, *H.influenzae*, *N.meningitidis MC58* (serogroup B), *C.jejuni*, *M.loti*, *C.crescentus*, *B.subtilis*, *B.halodurans*, *S.aureus N315* (MRSA), *C.acetobutylicum*, *M.genitalium*, *M.tuberculosis H37Rv* (lab strain), *C.pneumoniae CWL029*, *D.radiodurans*, *A.aeolicus*, *T.maritima*, *M.jannaschii*, *T.volcanium*, *S.solfataricus* y *A.pernix*.

Los análisis hechos con el índice de adaptación del codón (CAI) se realizaron con el programa CAI del EMBOSS, basado en el algoritmo construido por Sharp & Li (Sharp & Li 1987). La tabla de uso de codones del genoma se construyó con todos aquellos genes conservados que estuvieran presentes en *E.coli K-12 MG1655*, *E.coli K-12 W3110*, *E.coli O157 EDL933*, *E.coli O157 Sakai*, *E.coli CFT073*, *S.typhi CT18*, *S.typhimurium* y *Y.pestis CO92*. Los cálculos del CAI se hicieron para cada uno de los genes conservados y para los genes introducidos. Cuando el CAI de cualquiera de los genes introducidos estuvo a 4 desviaciones estándar (Ds) o más de la media de los genes conservados, el programa lo consideró como un gen de transferencia horizontal. La metodología del índice de GC3, sigue la misma lógica, excepto que aquí, se calculó el porcentaje de GC en la tercera posición del codón. Cuando el GC3 de un gen introducido estuvo a 4 Ds o más de la media de los genes conservados, el programa lo consideró como un gen de transferencia horizontal.

En la metodología de presencia-ausencia de un gen en los genomas cercanos (PD), se realizaron búsquedas de genes con Blastp (Altschul et al. 1997) de cada gen introducido en los genomas de *E.coli K-12 MG1655*, *E.coli O157 EDL933* y *S.typhimurium*, con un valor de e de 1×10^{-5} como umbral. Cuando el valor de e de un gen introducido resultó más alto que el umbral, el programa lo consideró como un gen de transferencia horizontal.

Se construyeron 5 modelos de Markov de primer a quinto orden. Cada uno de ellos se basa en distintas probabilidades; primer orden analiza dinucleótidos, segundo orden, trinucleótidos, y así sucesivamente. Los modelos de segundo o más órdenes no pudieron funcionar por falta de una semilla de datos suficientemente grande. Por ejemplo, para que un modelo de segundo grado funcionara se necesitarían genomas con varias decenas de miles de genes.

Para el modelo de Markov de primer orden (MM) se necesitaron dos diferentes grupos de probabilidades. El grupo "A" guarda las probabilidades de dinucleótidos de todos los genes del genoma, mientras que los grupos "B" tuvieron las probabilidades de dinucleótidos de cada uno de los genes introducidos. El grupo "A" se mantuvo constante durante todos los análisis, pero el grupo "B" se cambió cada vez que se analizó un gen introducido nuevo. Con estos datos, el índice de Markov se calcula de la siguiente forma:

$$S(m) = \frac{1}{h} \sum_{i=1}^{h-1} \frac{\log[P(xy)]_{set B_i}}{\log[P(xy)]_{set A}}$$

Donde $S(m)$ es el índice de Markov para la secuencia "m", h es el largo de la secuencia "m", $P(xy)$ set B_i son las probabilidades de dinucleótidos del gen introducido "i", y $P(xy)$ set A son las probabilidades de dinucleótidos de los genes conservados.

Se calculó una media de todos los índices de los genes conservados. Cuando el índice de un gen introducido estuvo a 4 o más Ds de esta media, el programa lo consideró como un gen transferido horizontalmente detectado.

Después del análisis se contaron, para cada método, el número de genes detectados. Éstos fueron separados de acuerdo al genoma de origen. El experimento se repitió exactamente 10 veces para cada método. También se realizaron los análisis para el CAI, el índice GC3 y el MM, usando todos los genes del genoma para los cálculos de los parámetros (media y desviación estándar) en lugar de sólo los conservados. La metodología de DP sólo pudo ser aplicada usando todos los genes de los genomas porque el método se construye sobre la base de búsqueda de genes en genomas completos. Se tomaron como genes de transferencia horizontal aquellos genes que estuvieron a 4 o más Ds.

Se analizaron los genomas de *E.coli* K12, *E.coli* O157 EDL933 y *S.typhimurium* usando el MM construido, buscando genes transferidos horizontalmente desde la divergencia de estas bacterias. Como grupo "A" se usaron todos los genes conservados, mientras que como grupos "B" todos aquellos genes no compartidos entre las enterobacterias, es decir, que estuvieran en *E.coli* pero que no estuvieran *S.typhimurium* ni en *Y.pestis* y que estuvieran en *S.typhimurium* pero no en *E.coli* y *Y.pestis*. Esto debido a que los genes transferidos horizontalmente, tras la divergencia de estos linajes, son más probables de encontrarse en este grupo de genes. Para evitar la detección de genes muy conservados recientemente duplicados, se removieron todos aquellos genes no-compartidos que tuvieran un 40% o más de similitud, a lo largo de toda la secuencia, con un gen conservado. Cuando uno de estos genes presentó un índice que estuviera a 4 o más Ds de la media de los genes conservados se tomó como un gen de transferencia horizontal reciente.

Se buscaron todos aquellos genes que pudieron haber perdido un poco de su arreglo original, tratando de dilucidar un proceso de mejoramiento (Lawrence & Ochman 1997). Para este experimento, se usó el MM para buscar aquellos genes que estuvieran entre 3 y 4 desviaciones estándar de la media de los genes conservados. Los nuevos genes detectados debieron pasar una selección para que pudieran ser validados como genes de transferencia horizontal: Debieron encontrarse adyacentes a un gen detectado a 4 o más Ds y haber presentado su misma distribución de ausencia-presencia en los genomas de *E.coli* K-12 MG1655, *E.coli* K-12 W3110, *E.coli* O157 EDL933, *E.coli* O157 Sakai, *E.coli* CFT073, *S.typhi* CT18, *S.typhimurium* y *Y.pestis* CO92.

Los genes de transferencia horizontal fueron clasificados de acuerdo a las rutas metabólicas

estipuladas en el KEGG. Se estudió caso por caso las anotaciones de cada uno de los genes. De esta forma se obtuvieron aquellas proteínas asociadas a membrana, enzimas, etc. Se hizo principal hincapié en los genes de profago y sus genes asociados, esto con el fin de definir el número de eventos y que genes se vieron involucrados (se denominaron islas de profagos). Se hizo un análisis de los genes aledaños a cada una de las islas para determinar la naturaleza de los genes que permiten la inserción de este tipo de elementos.

Se realizó una X^2 de esperados contra observados de la distribución de los GTH en las diferentes rutas metabólicas. Los esperados se calcularon: $[\text{\#total de GTH en el genoma}] * [\text{\#total de genes en cada vía metabólica}] / [\text{\#total de genes en el genoma}]$. Los valores significativos tuvieron que ser $p < 0.001$.

Se analizó también, usando búsquedas con Blastp (Altschul et al. 1997), qué genes compartían las dos *E.coli* y qué genes eran únicos de cada cepa. Así como el número de genes de transferencia horizontal que están duplicados dentro de cada genoma.

Duplicación de Fragmentos de DNA. La búsqueda de genes duplicados se realizó en *E.coli K-12 MG1655*, *E. coli O157 EDL933* y *S.typhimurium*. El programa Blastp (Altschul et al. 1997) se utilizó para realizar las búsquedas de cada gen dentro de su propio genoma con un umbral de 0.000001. A partir de las alineaciones realizadas por Blastp (Altschul et al. 1997) se corroboró que las proteínas fueran semejantes a lo largo de toda la secuencias.

A la lista de genes con por lo menos un parólogo, se clasificó y se les asignó función metabólica. Se estudió caso por caso las anotaciones de cada uno de los genes. De esta forma se obtuvieron aquellas proteínas asociadas a membrana, enzimas, etc.

Se realizó una X^2 de esperados contra observados de la distribución de las genes, con por lo menos una duplicación, en las diferentes rutas metabólicas. Los esperados se calcularon: $[\text{\#total de genes con por lo menos una duplicación en el genoma}] * [\text{\#total de genes en cada vía metabólica}] / [\text{\#total de genes en el genoma}]$. Los valores significativos tuvieron que ser $p < 0.001$.

La ancestría de los genes con por lo menos una duplicación se analizó buscando la presencia de las parejas en los genomas de las bacterias: *E.coli K-12 MG1655*, *E.coli K-12 W3110*, *E.coli O157 EDL933*, *E.coli O157 Sakai*, *E.coli CFT073*, *S.typhi CT18*, *S.typhimurium* y *Y.pestis CO92*. Aquellos genes que aparecieron sólo en una de las *E.coli* o sólo en una *S.typhimurium*, se definieron como recientemente duplicados en ese linaje. Estos genes se analizaron aparte, se contabilizaron y se investigó su distribución en las diferentes rutas metabólicas. Para analizar parejas más ancestrales, se obtuvieron los ortólogos de cada uno de las parejas de parálogos en *E.coli* presentes también en la mayoría de las proteobacterias γ (en un 60% de ellas). A las parejas resultantes se les investigó su distribución funcional.

Se analizaron las parejas de genes duplicados con el fin de determinar si estaban ambos en una misma vía metabólica o si se encontraban desempeñando funciones en vías distintas. Los casos en donde las proteínas aparecieron en la misma vía metabólica se estudiaron visualmente analizando los diagramas de las rutas metabólicas de *E.coli K12* a través de la página de internet del KEGG (<http://www.genome.ad.jp/kegg/pathway.html>). Este mismo análisis se hizo para las parejas recientes, y

para aquellas que se duplicaron antes de la divergencia de las proteobacterias.

Pérdida de Material Genético. Aquellos genes no compartidos que no fueron resultado de duplicación reciente ni de llegada por transferencia horizontal fueron tomados como genes perdidos en las otras cepas. Se cuantificaron los genes y se procedió a hacer una asignación en las rutas metabólicas. Posteriormente, se analizó caso por caso las anotaciones de los genes y con esa información se obtuvo cuáles estaban asociados a membrana, eran enzimas, etc.

Genes sin Pasado Evidente y Pérdida en la Capacidad de Detección de GTH. Se tomaron 25 genes al azar del genoma de *E.coli* K12 a quienes les fue imposible encontrarles algún rastro de su origen. Se construyeron tres modelos probabilísticos, uno para secuencias simples, uno para duplicación de fragmentos y uno para genes de transferencia horizontal. La idea principal era hacer que los tres modelos compitieran para decidir quién lograba explicar mejor la historia de ese gen. En una primera ronda de modelado, al gen que se iba a analizar se le cambiaron todas las bases de su secuencia por cada una de las tres opciones alternativas. Por ejemplo, para un gen de 100 nucleótidos se generarían 300 secuencias distintas, cada una de ellas con un solo cambio. Posteriormente se seleccionó el cambio puntal que hacía a la secuencia más simple, una vez traducido el cambio a aminoácidos (bajando el índice de complejidad del SEG), la volviera más parecida a algún otro gen del mismo genoma de *E.coli* (aumentándose su similitud), o adquiriera un arreglo de las bases más alejado a la media del genoma (basado en los índices de Markov). Cada vuelta fue escogida una nueva mutación que fuera encaminando a la molécula hacia alguno de estos tres destinos. Finalmente, el modelo de secuencias simples se detenía cuando el 20% de la secuencia del gen era una secuencia simple en aminoácidos, el modelo de duplicaciones cuando el gen se parecía por lo menos en un 25% de similitud a otro dentro del mismo genoma y el modelo de transporte horizontal, cuando el índice de Markov del gen rebasaba las 4 desviaciones estándar con respecto a la media de los genes conservados. El modelo que requirió menos cambios para lograr su objetivo fue el considerado como la mejor opción para explicar el origen de ese gen.

Para el análisis de pérdida de información se procedió a tomar 20 genes al azar de los siguientes genomas *A.pernix*, *S.solfataricus*, *M.jannaschii*, *D.radiodurans*, *C.pneumoniae*, *M.tuberculosis*, *M.genitalium*, *C.acetobutylicum*, *S.aureus* N315, *B.subtilis*, *C.crescentus*, *M.loti*, *V.cholerae*, *Y.pestis* CO92 y *H.influenzae*. También se tomaron al azar de *A.aeolicus*, *A.pernix*, *D.radiodurans*, *S.aureus* N315, *B.subtilis*, *C.crescentus*, *C.jejuni*, *Y.pestis* CO92 y *H.influenzae* 10 genes relacionados a vía metabólicas que en las enterobacterias presentaron una gran aceptación de GTH y 10 genes relacionados con vía que no aceptaron nuevos GTH. Posteriormente, a todos estos genes se les aplicó el mismo proceso de cambio en sus secuencias, pero con la salvedad de que ahora se seleccionaron sólo aquellas mutaciones que permitían que el índice de Markov del gen bajara de las 4 desviaciones estándar con respecto a la media de los genes conservados. Se contabilizó el número de cambios requeridos por cada gen y se normalizaron al tamaño de la secuencia.

Todos los programas que aquí no se especifican fueron escritos en PERL (Apéndice 1).

RESULTADOS Y DISCUSIÓN

Secuencias Simples (SeSp)

*"El tiempo es precioso, pero la verdad es
más preciosa que el tiempo".
Disraeli.*

Las secuencias simples (SeSp) aparecen en el 13% de todos los genes de los genomas analizados (Tabla 1) y presentan un sesgo composicional que corresponde a los aminoácidos más comunes en las células procariontes: L, A, G, V, I y S.

Se realizó un primer análisis para determinar la clase de proteínas a las que se asocian las SeSp. Los resultados mostraron que las SeSp tienen una predilección hacia las proteínas que están asociadas a las membranas (44% de los casos; Tabla 1 y 2). Como un acercamiento al papel que desempeñan las SeSp en las proteínas asociadas a membrana, se analizaron ubicaciones de las secuencias en las estructuras cristalográficas disponibles. Los resultados de la búsqueda de proteínas con SeSp y estructura cristalográfica en *E.coli* fueron dos: la subunidad III de la ubiquinol oxidasa y la subunidad c de la ATP sintetasa. En ambos casos las SeSp aparecen formando parte de una alfa hélice en las regiones intermembranales (Tabla 6). Tener dos resultados cristalográficos equivale a poco menos que nada para construir hipótesis consistentes. Fue por ello que la búsqueda se extendió a todas aquellas proteínas con estructura 3D determinada en *E.coli* que presentaran SeSp. Se obtuvieron ocho: las subunidades L1, L7/L12, L18, L20 y L23 del ribosoma, el factor σ 70 de la RNA polimerasa y las subunidades δ y δ' de la polimerasa III. En ellas las SeSp se ubican formando parte de asas, aunque aparecen también en alfas hélices (Tabla 6). Nuevamente el problema es que con diez datos cristalográficos es imposible elaborar conclusiones sobre la estructura que adoptan las regiones con SeSp en las proteínas.

Con la finalidad de obtener una mayor cantidad de información al respecto, se procedió a estudiar la presencia de las SeSp en los dominios intermembranales (de todas las proteínas de membrana) y la predicción de su estructura secundaria (de 25 proteínas de membrana con SeSp seleccionadas al azar; ver Métodos). Las secuencias simples aparecen en alfas hélices (45%), láminas beta (14%) y asas (41%). En la mitad de las proteínas de membrana (55.85%, 105 proteínas) las SeSp aparecieron dentro del dominio intermembranal. Están compuestas mayoritariamente por aminoácidos hidrofóbicos (el 68% de todos los aminoácidos en estas SeSp son hidrofóbicos). Estos resultados sugieren que las SeSp que están presentes en las proteínas de membrana, aparecen formando alfas hélices o asas mayoritariamente hidrofóbicas, que no se encuentran necesariamente inmersas en la membrana lipídica.

La segunda categoría funcional con mayor presencia de SeSp son las enzimas citosólicas (25%; Tabla 2): aparecen en ligasas, oxidoreductasas, cinasas y proteasas; aunque no es posible generalizar, ya que hay una enorme cantidad de este tipo de enzimas sin SeSp. El análisis de predicción de estructura secundaria (de 25 enzimas tomadas al azar; ver Métodos) mostró que las SeSp aparecen en asas (43%),

alfas hélices (40%) y láminas beta (17%). Además, están igualmente compuestas por aminoácidos hidrofóbicos (56%) e hidrofílicos (54%). Estos datos muestran nuevamente que las SeSp están preferentemente en alfas o en asas. En este caso, y debido a la naturaleza de las proteínas, la proporción de aminoácidos hidrofóbicos disminuye, lo cual indica que muchas de estas secuencias se encuentran en contacto con el medio acuoso.

Finalmente, la tercera categoría funcional con mayor presencia de SeSp es la de las proteínas hipotéticas (25%; Tabla 2). Para definir cuáles podrían corresponder a proteínas asociadas a membrana, se analizaron buscando dominios intermembranales. El 45.1% (122 casos) presentaron dichos dominios (se podrá suponer correctamente que el restante 54.9% serán enzimas citosólicas). Contrario a lo encontrado en las proteínas de membrana bien anotadas, las SeSp aparecieron en el 77.86% (95 casos) dentro de esos dominios. La predicción de la estructura secundaria (de 25 proteínas hipotéticas con SeSp tomadas al azar; ver Métodos), mostró los siguientes resultados: aparecen casi por igual en asas (44%) y en alfa hélices (40%), y son poco menos frecuentes en las láminas beta (16%). Además muestran una preferencia por los aminoácidos hidrofóbicos (61%).

Organismo	Núm. Genes del organismo que codifican para una proteína	Núm. Secuencias simples	SeSp asociadas a membrana	SeSp No asociadas a membrana	SeSp en genes parálogos
<i>E. coli K 12</i>	4287	549 (13 %)	199 (36 %)	151 (27 %)	180 (32.7 %)
<i>E. coli O157</i>	5348	638 (12 %)	240 (37 %)	128 (20 %)	246 (38.5 %)
<i>S. typhimurium</i>	4553	629 (14 %)	300 (47 %)	329 (47 %)	235 (37.3 %)

Tabla 1. Datos generales de las secuencias simples en las enterobacterias analizadas. O es *E. coli O157*, K es *E. coli K 12* y S es *S. typhimurium*. De las clasificaciones de proteínas con secuencias simples asociadas y no-asociadas a membrana fueron eliminados los genes hipotéticos.

Función asociada al gen	Presencia
Membrana	240(O), 199(K), 300(S)
Hipotéticos	112 (O), 163(K), 23(S)
Enzima	113(O), 97(K), 163(S)
Profago	72(O), 1(K), 20(S)

Tabla 2. Datos generales de las funciones asociadas a los genes con SeSp.

La distribución de las secuencias simples en las diferentes categorías metabólicas fue distinta a la esperada al azar ($p < 0.0001$). Siendo las categorías de proteínas asociadas a membrana, clasificación y degradación y movilidad celular las que tienen una mayor presencia parcial (Tabla 3). Todas estas categorías se caracterizan por tener un gran número de proteínas asociadas a membrana.

Ruta	<i>E. coli</i> K12		<i>E. coli</i> O157		<i>S. typhimurium</i>	
	Frecuencia de SeSp	% en esa ruta	Frecuencia de SeSp	% en esa ruta	Frecuencia de SeSp	% en esa ruta
M. de carbohidratos	31	11.90	29	10.58	41	13.80
M. energía	24	13.87	24	13.79	34	17.71
M. lípidos	8	16.00	7	14.58	8	16.33
M. nucleótidos	11	9.32	11	9.32	10	8.77
M. aminoácidos	29	11.65	28	11.43	38	14.56
M. de otros aminoácidos	9	12.33	7	10.29	8	11.43
M. de carbohidratos complejos	8	7.77	8	8.16	13	13.00
M. de lípidos complejos	10	15.87	11	18.00	12	18.46
M. de cofactores y vitaminas	10	5.78	9	5.00	18	9.57
Metabolitos secundarios	1	4.00	1	4.76	2	9.00
Biodegradación de xenobióticos	6	10.34	5	8.93	8	15.00
Transcripción	6	7.79	6	8.70	6	8.96
Traducción	11	8.46	11	8.40	10	7.69
Clasificación y degradación	18	32.14	16	26.23	22	26.51
Replicación y reparación	12	12.90	12	12.90	13	14.13
Transporte de membrana	106	23.93	109	24.17	85	21.85
Transducción de señales	12	12.63	13	13.40	16	18.18
Movilidad celular	13	22.00	13	23.64	21	26.92
No-asignados	295	11.00	384	10.32	356	12.28

Tabla 3. Distribución de las SeSp en las distintas categorías metabólicas en las tres enterobacterias analizadas. Se sombrearon los datos con mayor frecuencia y porcentaje relativo de SeSp.

Conociendo el tipo de proteínas que presentan SeSp, su ubicación, su estructura, su composición y su distribución en las categorías metabólicas ¿podemos entender su función y por lo tanto el por qué de su selección y mantenimiento? La respuesta es negativa. Con esta enorme cantidad de datos aún es desconocida la función de las SeSp dentro de las proteínas. En la bibliografía los únicos que han propuesto una función clara son Bayliss *et al* (2004), quienes han intentado relacionar este tipo de elementos con la diversidad de la patogenicidad de los procariontes y la estabilización de proteínas de membrana. El análisis hecho en las enterobacterias muestra que no existe una relación clara entre las SeSp y los elementos patogénicos. Hay dos adhesinas y una proteína de secreción tipo III en *E. coli* O157 y siete proteínas relacionadas con el aparato de secreción en *Salmonella*. La X^2 determinó que la presencia de las SeSp, dado el número de proteínas relacionadas con la patogenicidad en estas bacterias (adesinas, proteínas del sistema de secreción III, toxinas, proteínas asociadas a virulencia, proteínas de fagos; Dobrindt y Reidl 2000), es igual a la esperada y por lo tanto inexistente. El que la distribución y presencia de las secuencias

simples en las categorías funcionales de la cepa no-patógena y la cepa patógena de *E.coli* sean significativamente iguales ($p < 0.0001$) confirma estos datos.

Una posible explicación a la función y mantenimiento de las SeSp en estos genomas podría estar dada por la frecuencia y ubicación dentro del mapa metabólico de las SeSp seleccionadas recientemente. Para corroborar esta hipótesis se analizó la distribución de los genes con SeSp entre las tres enterobacterias. Las secuencias se comparten en un 75% entre las tres enterobacterias y en un 86% entre las dos cepas de *E.coli* (Tabla 4). Las secuencias aparecen conservadas, con la misma firma, en todos los casos analizados (50 genes).

¿El 15% de diferencia entre las tres bacterias podría deberse a la selección *de novo* de SeSp? La respuesta es negativa. Las secuencias que aparecen en sólo uno o dos linajes deben esta distribución a tres fenómenos en particular: la pérdida de algún gen que codificaba para ellas, o la ganancia de un gen con SeSp a través de una duplicación o de la transferencia horizontal. Es decir, los tres mecanismos que hacen que varíen las cifras de secuencias simples compartidas nada tienen que ver con la selección *de novo* de dichas secuencias. Sólo la presencia de algunas SeSp podrían atribuirse a selección *de novo* reciente. Estos son los casos de:

1. En *E. coli* K12 (2 genes en total). b0001 (péptido líder del operón Thr) y b3672 (péptido líder del operón ilvbn).
2. En *E. coli* O157 (11 genes en total). Z0607 (proteína putativa transportador de aminoácidos), Z1070 (proteína hipotética), Z1756 (proteína hipotética), Z2195 (proteína hipotética), Z2214 (proteasa de Zinc), Z2784 (proteína hipotética), Z2787 (permeasa), Z3292 (función desconocida), Z3487 (componente putativo de unión a algún sistema de transporte ATP dependiente), Z5166 (péptido líder del operón ilvbn) y Z5314 (arisulfatasa).
3. En *S. typhimurium* (3 genes en total). STM0001 (péptido líder del operón Thr), STM0274A (invasol SirA) y STM3797 (péptido líder del operón ilvB).

La selección de nuevas SeSp se ha dado a través de duplicaciones de genes o por la llegada de genes que ya presentaban la secuencia simple en su estructura. Son realmente pocos los casos donde la propia SeSp es la que se ha seleccionado *de novo*. Además, estas secuencias carecen de una función evidente, exceptuando claro está, a los péptidos líder y algunas proteínas putativas de membrana.

Núm. de genes que se comparten en las tres Bacterias	Núm. de genes que se comparten en las <i>E.coli</i>
446 (que equivalen al 81% de SeSp en K, al 70% en O y al 71% en S)	508 (que equivalen al 92.5% de SeSp en K y al 80% en O)

Tabla 4. Ancestría de las SeSp. O es *E. coli* O157, K es *E. coli* K 12 y S es *S. typhimurium*.

La pregunta que entonces surge es: ¿cuándo aparecieron el 75% de las SeSp que se comparten entre las tres enterobacterias? Para determinar el momento de aparición de la SeSp es necesario analizar la conservación de la secuencia simple entre los distintos ortólogos. Es importante analizar genes que no sean fácilmente sujetos a transferencia horizontal, ya que de lo contrario se volvería muy difícil la determinación del momento de aparición de las SeSp. Por ello, se decidió analizar todas las SeSp que aparecen en las proteínas de *E. coli* con estructura 3D determinada. Es poco probable que dichas proteínas, presentes en la Tabla 5, se transfieran horizontalmente (por su elevada presión de selección), además de que están ampliamente repartidas a lo largo y ancho del árbol filogenético de las especies.

Función asociada al gen	Sesgo Composicional	Posición en Estructura	Conservación de la SeSp en los diferentes linajes
Proteínas ribosomales:			
L20	<i>R</i>	<i>asa</i>	<i>T</i>
L18	<i>R</i>	<i>asa (asociado al tRNA)</i>	<i>H</i>
L23	<i>V</i>	<i>asa</i>	<i>M-P</i>
L1	<i>AE</i>	<i>asa</i>	<i>T</i>
L7/L12	<i>AVE</i>	<i>asa</i>	<i>T</i>
Factores de iniciación de la traducción:			
IF-2	<i>E y G</i>	?	<i>M-P</i>
IF-3	<i>K</i>	?	<i>M-P</i>
Factor σ 70	<i>D</i>	<i>asa</i>	<i>H</i>
Factor σ 54 ó 60	<i>Q</i>	?	<i>M-P</i>
RNA polimerasa subunidad Ω	<i>P</i>	?	<i>H</i>
DNA polimerasa I	<i>A</i>	?	<i>H</i>
ATP sintetasa subunidad c	<i>G</i>	<i>alfa</i>	<i>H</i>
Citocromo:			
Citocromo oxidasa putativa	<i>A</i>	?	<i>M-P</i>
Citocromo b(561)	<i>L</i>	?	<i>M-P</i>
Ubiquinol oxidasa subunidad III	<i>H</i>	<i>alfa-asa</i>	<i>M-P</i>
tRNA sintetetasas:			
Valine	<i>L</i>	?	<i>T</i>
Glicina sub β	<i>A</i>	?	<i>H</i>
Cisteína	<i>A</i>	?	<i>T</i>

Función asociada al gen	Sesgo Composicional	Posición en Estructura	Conservación de la SeSp en los diferentes linajes
Polimerasa III subunidad τ - γ	<i>P</i>	?	<i>H</i>
Polimerasa III δ	<i>L</i>	<i>alfa</i>	<i>T</i>
Polimerasa III δ'	<i>A</i>	<i>alfa-asa-alfa</i>	<i>T</i>
ATP sintetasa proteína I	<i>V</i>	?	<i>H</i>

Tabla 5. Proteínas con SeSp y estructura cristalográfica determinada. Se señala el sesgo composicional y el tipo de estructura secundaria en la que aparece la región simple, así como la conservación de la firma en todos los linajes (T), a partir de la proteobacterias γ o mal conservada (M-P) o el simple hecho de presentar una SeSp en una región determinada de su estructura (H).

Cada una de las proteínas de la Tabla 5 fue alineada con todos sus ortólogos para obtener el perfil de conservación de las SeSp. Sólo las SeSp que se encuentran en las subunidades L1, L7 y L20 del ribosoma (que están cerca del sitio catalítico del ribosoma; Yusupov 2001), las subunidades δ y δ' de la polimerasa III y las tRNA sintetasa de valina y cisteína, están bien conservadas en todos los linajes bacterianos (se mantuvieron conservadas entre un 90% y un 50%).

Es posible que la firma simple se hubiera perdido en las demás proteínas, pero que estuviera seleccionado el hecho de presentar una SeSp en cierta región específica. Esto resultó cierto para la subunidad L18 del ribosoma (alejada del sitio catalítico; Yusupov 2001) la subunidad c y la proteína I de la ATP sintetasa, la subunidad τ - γ de la polimerasa III, el factor σ 70, la tRNA sintetasa de glicina y la subunidad Ω de la RNA polimerasa.

Finalmente, en la subunidad L23 del ribosoma (alejada del sitio catalítico; Yusupov 2001), en la polimerasa I, en el factor σ 54, en la citocromo oxidasa putativa, la citocromo b(561), la subunidad III de la ubiquinol oxidasa, IF-2 e IF-3, las SeSp se presentaron en zonas de muy alto cambio, y sólo se logró apreciar cierta conservación de las firmas entre las proteobacterias. Además, la presencia de SeSp en una cierta región de la proteína se da sólo en algunos organismos repartidos a lo largo y ancho del árbol filogenético.

Algunas de las SeSp que se detectan en la actualidad parecen ser muy antiguas; aún si la firma no se ha mantenido íntegra, el hecho de presentarla en una región específica de la proteína sí lo ha estado. Otras SeSp aparecen a partir de la divergencia de las proteobacterias y son realmente pocos los casos en donde la SeSp está presente sólo en las enterobacterias. La pregunta que surge es ¿qué significado tiene que muchas de las SeSp que se detectan en la actualidad tuvieron un origen que se remonta tan atrás en el tiempo?

Para abordar el problema se realizó un examen de la cobertura de las secuencias simples en sus respectivos genes. Cuando el porcentaje de cobertura de las SeSp es mayor a un 20%, por lo general, las proteínas son desconocidas e hipotéticas o son péptidos líder de algún operón (operón de trh e ilvGEP). En la cepa patógena de *E.coli*, 11 de estas 26 proteínas pertenecían a profagos, lo mismo que tres proteínas de *S. typhimurium*. Son particularmente interesantes los casos de una proteína de choque ácido con 58% de SeSp ricas en A y T, la proteína de membrana STM4222 que presentó SeSp ricas en serinas en un 56% de su secuencia, la subunidad c de la ATP sintetasa con un 20.5% de SeSp ricas en A y G y la subunidad L1

del ribosoma con un 30.2% de SeSp ricas en A y E.

Las proteínas con SeSp sumamente antiguas que además presentan un alto porcentaje de cobertura, podrían ser los remanentes de un mundo antiguo donde las proteínas eran más sencillas.

Transferencia Horizontal

*"Una cosa no es necesariamente cierta porque
un hombre muera por ella".
Oscar Wilde.*

La transferencia horizontal es un fenómeno biológico cuya comprensión detallada que ha eludido a los investigadores por mucho tiempo. Por esta razón, ha sido muy difícil determinar su importancia, su frecuencia, su intensidad y los genes que involucra. Esta investigación requirió identificar a los genes de transferencia horizontal presentes en las enterobacterias, con el fin de determinar su participación en la estructuración de las rutas metabólicas. Como primera aproximación se realizó una búsqueda bibliográfica de todos aquellos métodos desarrollados hasta ese momento (Aravind et al. 1998; Hayes & Borodovsky 1998; Kyrpidis et al. 1999; Ochman & Lawrence 1998; Garcia-Vallvé et al. 2000; Koonin 2001; Ragan y Charlebois 2002; Hooper & Berg 2002; Snel et al. 2002; Daubin, Moran & Ochman 2003). Los más utilizados resultaron ser: A) El índice de adaptación del codón (CAI), que parte de la base que los genes transferidos horizontalmente tendrán una menor tasa de transcripción porque no están adecuados al uso de codones del organismo. Así, aquellos genes con un CAI muy bajo podrían ser genes transferidos horizontalmente (Ochman & Lawrence 1998). Las comparaciones se hacen para cada gen contra un promedio de CAI general de todo el genoma. B) El porcentaje de GC en la tercera posición del codón (GC3); en este caso se calculan los porcentajes para cada gen y se compara contra la media de GC3 de todo el genoma (Garcia-Vallvé et al. 2000). C) Modelos de Markov de quinto orden (MM5); inicialmente éstos se aplicaron en la detección de ORF en genomas no anotados. Los modelos se alimentaban con todos los ORF conocidos y generaban un modelo típico de quinto orden, es decir, un conjunto de probabilidades de hexanucleótidos. Aquellos genes que eran mejor explicados por un modelo atípico se etiquetaban como genes de transferencia horizontal (Hayes & Borodovsky 1998). D) Los análisis de distribución de genes (PD) en organismos cercanos filogenéticamente; cuando un gen se encuentra presente en un genoma pero ausente en los organismos cercanos, dados ciertos umbrales estrictos, es indicio de que el gen fue transferido horizontalmente (Ragan y Charlebois 2002; Hooper y Berg 2002; Snel et al. 2002; Daubin, Moran y Ochman 2003).

Ragan (2001) realizó un trabajo trascendental en lo que a transferencia horizontal se refiere. En él, aplicó cuatro métodos diferentes de detección de genes de transferencia horizontal en el genoma de *E.coli* (CAI, MM5, PD y DP; este último se basa en la comparación de promedios de valores de Blastp entre dos genomas, los genes que están lejos de la media de estos promedios pueden ser genes de transferencia horizontal; Clarke et al. 2002). Se encontró que cada uno de ellos detectaba un grupo de genes distintos.

Marc Ragan concluyó que probablemente los resultados se debían a que los métodos composicionales detectaban genes transferidos recientemente y los métodos filogenéticos detectaban genes más antiguos. ¿Es suficiente esta explicación?

El primer experimento que se realizó fue la repetición del ejercicio que hizo Mark Ragan pero usando solamente los modelos más comunes (CAI, GC3 y PD) en el genoma de *E.coli*. Se corroboró que cada uno de los métodos detectaba un grupo de genes distinto, con pocos o ningún gen compartido. Pese a que los modelos tenían bases robustas, los resultados no los parecían apoyar. Una pregunta lógica que se desprendió de estos primeros resultados fue: ¿Están los modelos realmente detectando transferencia horizontal?

Detectando genes introducidos artificialmente. El principal problema al que se ven enfrentados los modelos de detección, es que no es posible saber que genes tuvieron un origen horizontal de antemano. Por ello y para corroborar la efectividad de los métodos es necesario diseñar un control experimental, ajeno a los genes que componen un genoma, que permita determinar objetivamente si se están detectando genes transferidos horizontalmente (GTH). Se puede realizar una simple prueba teórica: la inclusión artificial de genes conocidos provenientes de otros organismos dentro de cualquier otro genoma que sirva de modelo experimental. Todos los métodos que realmente estén funcionando no tendrían ningún problema en detectar a los genes introducidos artificialmente.

Para desarrollar esta la idea, se eligió al genoma de *E.coli* como modelo (una cepa patógena O157 y una no-patógena K12) y se decidió evaluar los métodos de GC3, CAI y PD. Se escogieron 21 organismos distintos repartidos por todo el árbol filogenético (con el fin de analizar la capacidad de detección de los métodos a través de la similitud entre sus genomas; ver Métodos), y se añadieron 100 genes provenientes de cada uno de ellos a los genomas modelo (ver Métodos).

Todos los métodos mostraron limitaciones severas en la detección de los genes exógenos (Tabla 6). Ninguno de las dos estrategias composicionales (GC3 y CAI) detectó más del 20% de los genes añadidos (Tabla 6). CAI tuvo los mejores resultados cuando los genes eran de arqueas y GC3, cuando venían de genomas cuyo promedio de GC es mucho menor o mayor que el de *E.coli* (50.2% de GC genómico): fueron los casos de *M.genitalium* (con 31.7% de GC) y *C.jejuni* (con 30.5% de GC). Estos datos confirman las fallas en detectar GTH por estos dos métodos (Koski et al. 2001; Wang 2001), las cuales se deben a la heterogeneidad composicional que está presente en cualquier genoma, particularmente la relativa a los índices de CAI y GC3 (Daubin & Perrière 2003).

PD pareció funcionar bien, ya que detectaba alrededor de un 50% de los genes introducidos artificialmente (Tabla 6). El problema es que la única detección que superó el 80% fue con los genes provenientes de la arquea, *A.pernix*. El modelo perdía capacidad de detección cuando se trataba de genes de genomas muy cercanos filogenéticamente, en donde las detecciones bajaron drásticamente hasta un 25%. Los fallos en esta metodología ampliamente utilizada (Ragan & Charlebois 2002; Daubin et al. 2003a; Daubin et al. 2003b), prueban que aún con un umbral estricto ($e < 0.00001$) se pierde la detección de

muchos de los genes transferidos, probablemente por la presencia de ortólogos o por falsas parejas de homólogos. Aunque este enfoque parecía funcionar mejor que los dos métodos de análisis de composición probados, no era lo suficientemente bueno exacto para los propósitos del trabajo. Era necesario un modelo que tuviera una detección de por lo menos un 80% de los GTH añadidos artificialmente. Hasta ese momento ninguno cumplía con el requisito.

	<i>Escherichia coli</i> O157 EDL			<i>Escherichia coli</i> K12 MG1655		
	GC3	CAI	PD	GC3	CAI	PD
<i>A.pemix</i>	0.5 ± 0.5	11.33 ± 3.1	82.3 ± 0.7	0.15 ± 0.15	39.6 ± 3.4	82.3 ± 0.7
<i>S.solfataricus</i>	0.3 ± 0.3	4.5 ± 2.3	71.4 ± 2.8	1.6 ± 1.1	27.1 ± 3.5	70.5 ± 3
<i>T.volcanium</i>	0	1.17 ± 0.9	62.4 ± 4.6	0.5 ± 0.5	10.1 ± 4	62.6 ± 4.1
<i>M.jannaschii</i>	0.5 ± 0.5	0.7 ± 0.4	70.2 ± 4.6	17.8 ± 6.2	4.2 ± 1.9	70 ± 4.4
<i>T.maritima</i>	0	0.7 ± 0.7	42.5 ± 4.5	0	0.9 ± 0.9	42.7 ± 4.9
<i>A.aeolicus</i>	0	2.7 ± 2.1	38.9 ± 3.7	0.5 ± 0.5	15.7 ± 2.6	38.4 ± 3.5
<i>D.radiodurans</i>	6.5 ± 2.5	0	58.8 ± 2.7	20.6 ± 3.6	0	59.1 ± 2.8
<i>C.pneumoniae</i>	0	0.5 ± 0.5	54.3 ± 6	0.7 ± 0.7	1.3 ± 0.8	54.6 ± 6.1
<i>M.tuberculosis</i>	0.7 ± 0.7	0	61.6 ± 5.3	1.9 ± 1.6	0	61.6 ± 5.4
<i>M.genitalium</i>	11.83 ± 2.9	0	44.2 ± 5.8	38.0 ± 4.1	0	44.6 ± 5.7
<i>C.acetobutylicum</i>	9.67 ± 1.8	0.4 ± 0.4	54.5 ± 4.1	46.8 ± 4.5	5 ± 1.7	54.9 ± 4.3
<i>S.aureus_n315</i>	4 ± 2.9	0	50.4 ± 4	33.7 ± 4.5	0.4 ± 0.4	49.8 ± 4.5
<i>B.halodurans</i>	0	0	49.8 ± 7.2	0	0	49.3 ± 6.8
<i>B.subtilis</i>	0	0	48.7 ± 3.9	0.9 ± 0.6	0	48.3 ± 4.1
<i>C.crescentus</i>	21.5 ± 5.9	0	51.8 ± 4.4	44.6 ± 5.6	0.15 ± 0.15	51.9 ± 3.7
<i>M.loti</i>	0.6 ± 0.6	0	49.8 ± 4.9	2.1 ± 1.5	0.15 ± 0.15	50.8 ± 4.8
<i>C.jejuni</i>	19.67 ± 5.1	0	40.9 ± 3.7	69.3 ± 2.5	0	40.8 ± 4.9
<i>N.meningitidis</i>	0.2 ± 0.2	0	46.5 ± 4.4	2.4 ± 1.1	0.5 ± 0.5	46.8 ± 4.2
<i>Y.pestis</i>	0	0	26.2 ± 5.1	0	0	29.6 ± 4.9
<i>H.influenzae</i>	0.8 ± 0.7	0	21.5 ± 4	5.7 ± 2.1	0	20.7 ± 3.9
<i>S.typhimurium LT2</i>	0	0	23.5 ± 3.8	0.1 ± 0.1	0	26.7 ± 4.3

Tabla 6. Porcentaje de detección de los genes introducidos con los métodos de GC3, CAI y PD, usando todos los genes del genoma para calcular los parámetros. La detección máxima es de 100 porque se introdujo esa cantidad de genes en los genomas de *E.coli* provenientes de cada uno de los genomas que se enlistan. Se sombreadon los datos con una detección superior al 80% de lo genes artificialmente introducidos.

La opción que faltaba de tomar en cuenta eran los modelos de Markov. Se construyeron cinco modelos distintos, del primer al quinto orden. A los modelos de orden igual o superior a dos fue imposible evaluarlos porque los genes que hay en el genoma de *E.coli* no alcanzaban para el cálculo de probabilidades; se necesitarían genomas con millones de genes para poder alimentarlos. Por esta razón, Hayes y Borodovsky (1998) construyeron su MM5 utilizando todos los ORFs conocidos. Sin embargo, este

enfoque no es lo más adecuado cuando el objetivo es detectar genes de transferencia horizontal, ya que mezcla las probabilidades de genes de múltiples genomas. Por lo tanto, la opción viable fue el modelo de Markov de primer orden (MM). Los resultados del experimento control fueron alentadores (Tabla 7). En varios genomas, sobre todo los más alejados filogenéticamente, se tuvieron detecciones de hasta un 99.9%.

El modelo parecía funcionar aparentemente mejor en la cepa no-patógena que en la patógena; se estaba cometiendo un error teórico gigantesco: al hacer el análisis del genoma completo se estaban mezclando en los cálculos de probabilidades genes transferidos y genes no-transferidos (ver Métodos). De esta forma, los índices que se calculaban estaban equivocados, ya que tenían una media del genoma errónea y una desviación estándar más grande.

	<i>Escherichia coli</i> O157	<i>Escherichia coli</i> K12
	MM	MM
<i>A.pemix</i>	98.8 ± 1.2	99.9 ± 0.1
<i>S.solfataricus</i>	59.5 ± 3	92.1 ± 1.8
<i>T.volcanium</i>	18 ± 3.9	37.4 ± 4.5
<i>M.jannaschii</i>	75.5 ± 2.6	97.2 ± 1.6
<i>T.maritima</i>	78.5 ± 2	91.4 ± 3
<i>A.aeolicus</i>	78.83 ± 2	94.6 ± 2.6
<i>D.radiodurans</i>	16 ± 5	38.6 ± 5.1
<i>C.pneumoniae</i>	24.33 ± 4.6	48 ± 2.7
<i>M.tuberculosis</i>	10.5 ± 4	19.9 ± 2.8
<i>M.genitalium</i>	45.67 ± 2.3	47.1 ± 3.8
<i>C.acetobutylicum</i>	68.67 ± 4	94 ± 2.2
<i>S.aureus_n315</i>	29.5 ± 5.8	64.1 ± 2.3
<i>B.halodurans</i>	8.67 ± 2.4	17.9 ± 4.1
<i>B.subtilis</i>	8.5 ± 4.1	21.1 ± 4.5
<i>C.crescentus</i>	19.5 ± 4.7	51 ± 4.4
<i>M.loti</i>	16.67 ± 3.2	30.3 ± 4.4
<i>C.jejuni</i>	74.83 ± 3.3	91.3 ± 2.2
<i>N.meningitidis</i>	40.67 ± 4.5	52.7 ± 4.6
<i>Y.pestis</i>	7.83 ± 1.6	13.1 ± 3.3
<i>H.influenzae</i>	12.67 ± 2.9	31.2 ± 3.1
<i>S.typhimurium LT2</i>	7.17 ± 3.9	14.1 ± 2.4

Tabla 7. Porcentaje de detección de los genes introducidos con el modelo de Markov para las dos *E.coli* utilizadas. La detección máxima es de 100 porque se introdujo esa cantidad de genes en los genomas de *E.coli* provenientes de cada uno de los genomas que se enlistan. Se sombrearon los datos con una detección superior al 80% de los genes artificialmente introducidos.

El siguiente paso fue separar a los genes de estas dos *E.coli* en: claramente conservados y posiblemente transferidos. El modelo se alimentaría sólo de los genes conservados, eliminándose del cálculo de probabilidades todos aquellos genes de origen dudoso. Se probó el experimento control nuevamente. Los resultados fueron notables (Tabla 8): en casi todos los genomas, excepto los *Bacillus*, *N.meningitidis*, *Y.pestis*, and *S.typhimurium LT2*, el modelo detectó un 80% o más de los genes introducidos.

Se aplicó la misma lógica para las metodologías CAI y GC3 con el fin de ver si existían mejoras en sus detecciones. Los parámetros (media y desviación estándar) se calcularon sólo con los genes

conservados. Los resultados mejoraron (Tabla 8), pero no lograron superar a los obtenidos con el MM. CAI obtuvo buenas detecciones, por arriba del 80%, con los genes provenientes de *A.pernix*, *S.solfataricus* y *A.aeolicus*; tres organismos muy lejanos a *E.coli*. GC3 obtuvo buenas detecciones (por arriba del 80% de genes introducidos artificialmente) cuando los genes provenían de *C.crescentus*; cuyo genoma tiene 67% de GC; *M.jannaschii*, con 31.3% de GC en su genoma; *M.genitalium*, con 31.7% de GC; *C.acetobutylicum*, con 30.9% de GC; *S.aureus*, con 32.8% de GC y *C.jejuni*, con 30.5% de GC.

	Escherichia coli O157 EDL			Escherichia coli K12 MG1655		
	MM	GC3	CAI	MM	GC3	CAI
<i>A.pernix</i>	100	5.7 ± 3	86.3 ± 3.4	100	4 ± 1.8	87.9 ± 2.1
<i>S.solfataricus</i>	100	49 ± 4.2	86.9 ± 3.7	99.9 ± 0.1	48.8 ± 3.5	90.4 ± 3.5
<i>T.volcanium</i>	97.7 ± 1.2	4.9 ± 3.2	60.2 ± 5.9	98.8 ± 0.9	4.1 ± 1.5	72.3 ± 3.9
<i>M.jannaschii</i>	100	91.3 ± 3.3	44.5 ± 4.8	100	90.7 ± 2.8	55.6 ± 5.6
<i>T.maritima</i>	99 ± 1	0.4 ± 0.4	22.4 ± 5.2	98.8 ± 0.9	0.5 ± 0.5	29.5 ± 4.2
<i>A.aeolicus</i>	99.8 ± 0.2	2.5 ± 0.9	83.6 ± 2.7	99.9 ± 0.1	2.2 ± 0.6	89.9 ± 4.2
<i>D.radiodurans</i>	96 ± 2.76	74.5 ± 4.6	0.7 ± 0.6	94.6 ± 1.8	71.9 ± 5.1	0.7 ± 0.6
<i>C.pneumoniae</i>	98.6 ± 1.2	16.8 ± 2.1	27.6 ± 4.6	98.8 ± 1	18.2 ± 2.5	25.9 ± 4.6
<i>M.tuberculosis</i>	90.1 ± 2.62	31.8 ± 4	0	92.3 ± 2.1	28.7 ± 4.9	0.1 ± 0.1
<i>M.genitalium</i>	99.5 ± 0.5	84.7 ± 3.6	2.1 ± 1	99.5 ± 0.5	83.8 ± 3.2	3.7 ± 1.4
<i>C.acetobutylicum</i>	100	98 ± 1.8	39.2 ± 4.6	99.8 ± 0.2	97.6 ± 1.5	48.9 ± 4.2
<i>S.aureus_n315</i>	99.7 ± 0.3	94.9 ± 2.3	4 ± 2.1	99.9 ± 0.1	95.6 ± 2.5	3.9 ± 1
<i>B.halodurans</i>	50.6 ± 3.41	2.7 ± 2.7	2.2 ± 1.2	50.4 ± 4	3.8 ± 1.5	2.8 ± 1.5
<i>B.subtilis</i>	57.3 ± 4	8.6 ± 2.5	1.2 ± 1.1	55.4 ± 3	8.1 ± 3.1	1.4 ± 1.1
<i>C.crescentus</i>	98.1 ± 1.3	79.8 ± 3	0.8 ± 0.7	99 ± 1	78.7 ± 3.7	0.7 ± 0.6
<i>M.loti</i>	89.8 ± 2.79	42.1 ± 2.2	0.4 ± 0.4	91.5 ± 4	39.6 ± 3.5	0.3 ± 0.3
<i>C.jejuni</i>	100	98.1 ± 1.1	3.2 ± 1.5	100	99.1 ± 0.8	3.8 ± 1
<i>N.meningitidis</i>	70.7 ± 4.9	5.3 ± 2.1	2 ± 1.5	71.8 ± 4	5.8 ± 2	2.2 ± 1.6
<i>Y.pestis</i>	27.1 ± 3.5	1.8 ± 1.4	0.9 ± 0.7	27.5 ± 4.3	3.4 ± 1.9	0.9 ± 0.8
<i>H.influenzae</i>	85.7 ± 3.2	60.6 ± 5.3	0.7 ± 0.4	86.2 ± 3.8	61 ± 4	1.1 ± 1.1
<i>S.typhimurium LT2</i>	24.3 ± 3.6	1.6 ± 1	0.4 ± 0.4	25.1 ± 3.6	1.1 ± 1.1	0.6 ± 0.6

Tabla 8. Porcentaje de detección de los genes introducidos con los métodos de GC3, CAI y MM, usando sólo los genes conservados. La detección máxima es de 100 porque se introdujo esa cantidad de genes en los genomas de *E.coli* provenientes de cada uno de los genomas que se enlistan. Se sombrearon los datos con una detección superior al 80% de lo genes artificialmente introducidos.

La idea de utilizar a los genes conservados como semilla para el cálculo de los parámetros no pudo aplicarse a la metodología PD porque el método se basa en la comparación total de genes entre genomas. Aunque el método de GC3 pareció funcionar mucho mejor que el de CAI, no existe en los resultados una coherencia filogenético: el método funciona muy bien sólo detectando aquellos genes que provienen de genomas cuyo porcentaje de GC es muy diferente al del genoma que se quiere analizar, en este caso *E.coli* (50.2% de GC genómico).

Aplicación del Modelo de Markov para detectar los genes recientemente transferidos en el genoma de *E. coli* K12 MG1655, *E. coli* O157 EDL933 y *S. typhimurium* LT2. En vista de que el modelo de Markov resultó ser la estrategia experimental más adecuada, se decidió aplicarla para buscar aquellos genes que pudieran haber llegado a partir de la divergencia entre *E. coli* y *S. typhimurium* LT2. Se usaron los genes conservados para calcular los parámetros y se buscaron los GTH en el conjunto de genes que no están compartidos entre estas bacterias. La aplicación del modelo fue llevada a cabo con el objetivo de distinguir entre pérdida de genes y transferencia horizontal (Tabla 9).

Organismo	Total de GTH propuestos	A 4 Ds o más	Entre 3 y 4 Ds	Genes en islas de fagos	Genes que no están en islas de fagos	Genes perdidos en los otros linajes	Genes aparentemente duplicados
<i>E. coli</i> K12	429	315 (73.5%)	114 (26.5%)	22 (5%) en 3 eventos	407 (95%)	450	93
<i>E. coli</i> O157 EDL933	1127	831 (74%)	296 (26%)	658 (58%) en 26 eventos	469 (52%)	584	499
<i>S. typhimurium</i> LT2	510	367 (72%)	143 (28%)	127 (25%) en 6 eventos	383 (75%)	352	96

Tabla 9. Número de genes recientes transferidos horizontalmente detectados por el MM en el genoma de *E. coli* K12, *E. coli* O157 EDL933 y *S. typhimurium*. El número total de genes encontrados está dividido en aquellos a cuatro o más desviaciones estándar (Ds) y aquellos entre tres y cuatro desviaciones estándar (Ds), pero adyacentes a un gen previamente detectado a cuatro o más Ds (ver Métodos). Se muestran los genes según su relación con islas de profago y el número de estos genes que pudieran tener, al menos, un gen parálogo. Los genes no-compartidos que no fueron detectados por el modelo de Markov fueron considerados como genes perdidos en los otros linajes.

Cuando se aplicó el modelo a los genomas de *E. coli* K12 MG1655, *E. coli* O157 EDL933 y *S. typhimurium* se hizo una primera búsqueda de todos aquellos genes que estuvieran a cuatro o más desviaciones estándar de la media de los genes conservados. Se encontraron muchos genes transferidos recientemente, particularmente en *E. coli* O157 EDL933 (Tabla 9) donde la mayoría de los genes llegaron a través de fagos, como ya había sido mostrado por Perna *et al.* (Perna *et al.* 2001). *S. typhimurium* tuvo menor llegada de genes por fagos, (McClell *& et al.* 2001), mientras que la cepa no-patogénica de *E. coli* presentó una casi inexistente llegada de genes por este proceso (Blattner *et al.* 1997).

Se hizo un análisis de los genes aledaños a cada una de las islas, para determinar la naturaleza de los genes que permiten la inserción de GTH. Se encontró una infinidad de genes colindantes distintos, lo que indica que estos elementos se incorporan al genoma de manera aleatoria.

Debido a que el modelo de Markov funciona adecuadamente cuando los genes conservan parte de su arreglo original (Tabla 8), su aplicación en un caso real podría implicar el no detectar algunos genes que han acumulado ciertas mutaciones que los hacen más parecidos al genoma del hospedero. Por lo tanto, se decidió hacer una segunda vuelta de detecciones en la cual se buscarían (en las tres enterobacterias modelo) todos aquellos genes que estuvieran entre tres y cuatro desviaciones estándar adyacentes a un genes detectado en la primera ronda de búsqueda y con su misma distribución (ausencia-presencia) en las enterobacterias. El número de genes en esta situación fue similar en las tres bacterias. Su frecuencia es

Aplicación del Modelo de Markov para detectar los genes recientemente transferidos en el genoma de *E. coli* K12 MG1655, *E. coli* O157 EDL933 y *S. typhimurium* LT2. En vista de que el modelo de Markov resultó ser la estrategia experimental más adecuada, se decidió aplicarla para buscar aquellos genes que pudieran haber llegado a partir de la divergencia entre *E. coli* y *S. typhimurium* LT2. Se usaron los genes conservados para calcular los parámetros y se buscaron los GTH en el conjunto de genes que no están compartidos entre estas bacterias. La aplicación del modelo fue llevada a cabo con el objetivo de distinguir entre pérdida de genes y transferencia horizontal (Tabla 9).

Organismo	Total de GTH propuestos	A 4 Ds o más	Entre 3 y 4 Ds	Genes en islas de fagos	Genes que no están en islas de fagos	Genes perdidos en los otros linajes	Genes aparentemente duplicados
<i>E. coli</i> K12	429	315 (73.5%)	114 (26.5%)	22 (5%) en 3 eventos	407 (95%)	450	93
<i>E. coli</i> O157 EDL933	1127	831 (74%)	296 (26%)	658 (58%) en 26 eventos	469 (52%)	584	499
<i>S. typhimurium</i> LT2	510	367 (72%)	143 (28%)	127 (25%) en 6 eventos	383 (75%)	352	96

Tabla 9. Número de genes recientes transferidos horizontalmente detectados por el MM en el genoma de *E. coli* K12, *E. coli* O157 EDL933 y *S. typhimurium*. El número total de genes encontrados está dividido en aquellos a cuatro o más desviaciones estándar (Ds) y aquellos entre tres y cuatro desviaciones estándar (Ds), pero adyacentes a un gen previamente detectado a cuatro o más Ds (ver Métodos). Se muestran los genes según su relación con islas de profago y el número de estos genes que pudieran tener, al menos, un gen parálogo. Los genes no-compartidos que no fueron detectados por el modelo de Markov fueron considerados como genes perdidos en los otros linajes.

Cuando se aplicó el modelo a los genomas de *E. coli* K12 MG1655, *E. coli* O157 EDL933 y *S. typhimurium* se hizo una primera búsqueda de todos aquellos genes que estuvieran a cuatro o más desviaciones estándar de la media de los genes conservados. Se encontraron muchos genes transferidos recientemente, particularmente en *E. coli* O157 EDL933 (Tabla 9) donde la mayoría de los genes llegaron a través de fagos, como ya había sido mostrado por Perna *et al.* (Perna *et al.* 2001). *S. typhimurium* tuvo menor llegada de genes por fagos, (McClell *et al.* 2001), mientras que la cepa no-patogénica de *E. coli* presentó una casi inexistente llegada de genes por este proceso (Blattner *et al.* 1997).

Se hizo un análisis de los genes aledaños a cada una de las islas, para determinar la naturaleza de los genes que permiten la inserción de GTH. Se encontró una infinidad de genes colindantes distintos, lo que indica que estos elementos se incorporan al genoma de manera aleatoria.

Debido a que el modelo de Markov funciona adecuadamente cuando los genes conservan parte de su arreglo original (Tabla 8), su aplicación en un caso real podría implicar el no detectar algunos genes que han acumulado ciertas mutaciones que los hacen más parecidos al genoma del hospedero. Por lo tanto, se decidió hacer una segunda vuelta de detecciones en la cual se buscarían (en las tres enterobacterias modelo) todos aquellos genes que estuvieran entre tres y cuatro desviaciones estándar adyacentes a un genes detectado en la primera ronda de búsqueda y con su misma distribución (ausencia-presencia) en las enterobacterias. El número de genes en esta situación fue similar en las tres bacterias. Su frecuencia es

mucho menor a los genes previamente detectados.

Es importante aclarar que muchos de los eventos de llegada podrían estar sobre-representados debido a la enorme cantidad de genes de transferencia horizontal que se han duplicado desde que llegaron a ese genoma (Tabla 9; Hooper y Berg 2003b), muchos de los cuales corresponden precisamente a las islas de profago. Hay 282 genes de estas islas con al menos una duplicación en el genoma de *E. coli* O157, y 107 en *S. typhimurium*. El problema es que es imposible definir si estos genes realmente se duplicaron o si simplemente hubo más de una copia del genoma del virus que se insertó. En ese caso cada evento sería independiente.

La X^2 de la distribución de los genes en las rutas metabólicas fue con una $p < 0.0001$ distinta a la esperada. Esto se refleja en los resultados ya que casi todos los genes transferidos horizontalmente están presentes en la categoría de no-asignados (Tabla 10), mientras que hay vías que no han tenido ninguna llegada de GTH seleccionada.

Ruta	<i>E. coli</i> K12		<i>E. coli</i> O157		<i>S. typhimurium</i>	
	Frecuencia de GTH	% en esa ruta	Frecuencia de GTH	% en esa ruta	Frecuencia de GTH	% en esa ruta
M. de carbohidratos	4	1.44	7	2.55	14	4.71
M. energía	0	0.00	0	0.00	4	2.00
M. lípidos	0	0.00	1	2.00	1	2.00
M. nucleótidos	0	0.00	0	0.00	0	0.00
M. aminoácidos	3	1.20	3	1.22	6	2.30
M. de otros aminoácidos	3	4.11	1	1.47	0	0.00
M. de carbohidratos complejos	3	2.91	1	1.00	2	2.00
M. de lípidos complejos	2	3.17	1	1.64	2	3.00
M. de cofactores y vitaminas	2	1.16	1	0.56	1	0.53
Metabolitos secundarios	0	0.00	0	0.00	0	0.00
Biodegradación de xenobióticos	2	3.45	1	1.79	0	0.00
Transcripción	1	1.30	0	0.00	0	0.00
Traducción	0	0.00	0	0.00	0	0.00
Clasificación y degradación	3	5.36	1	1.64	0	0.00
Replicación y reparación	1	1.00	1	1.00	0	0.00
Transporte de membrana	11	2.48	7	1.55	5	1.29
Transducción de señales	1	1.00	1	1.00	0	0.00
Movilidad celular	0	0.00	0	0.00	0	0.00
No-asignados	382	14.33	1088	29.25	345	11.90

Tabla 10. Distribución general de los genes transferidos horizontalmente en las distintas categorías metabólicas en las tres enterobacterias analizadas. Se sombreadon los datos con mayor frecuencia y porcentaje relativo de GTH.

Aunque el número de genes transferidos recientemente en las cepas patogénicas y no-patogénica es diferente, la naturaleza de las proteínas involucradas es muy similar. La gran mayoría de ellas pertenecen a la categoría de no-asignados (Tabla 13; Daubin, Moran y Ochman 2003). Pese a esto, un análisis más refinado muestra que muchas de las proteínas para las que codifican estos genes, tienen funciones muy bien definidas: Proteínas de profago, transposasas y elementos de inserción, proteínas de membrana y enzimas (Figura 4). En las cepas patogénicas, se encontraron proteínas que están directamente relacionadas con la patogenicidad (Dobrindt y Reidl 2000) como las adhesinas, las proteínas del sistema de secreción tipo III y algunas proteínas de virulencia (Figura 1). Los genes no-compartidos que permanecieron sin detectar por el modelo de Markov fueron tomados como pérdidas. El análisis funcional de estos genes muestra que codifican principalmente para proteínas de membrana o enzimas no-asignadas. Pero en algunos casos estos genes sí presentaron asignación y correspondían a funciones muy conservadas dentro de las células (Figura 4). Para validar la aplicación del modelo de Markov en un caso real, se contó el número de posibles falsos negativos (número de genes con funciones conservadas en el grupo de GTH) y falsos positivos (genes de profago no detectados por el modelo). Se encontraron pocos genes en ambas clases (Figura 4).

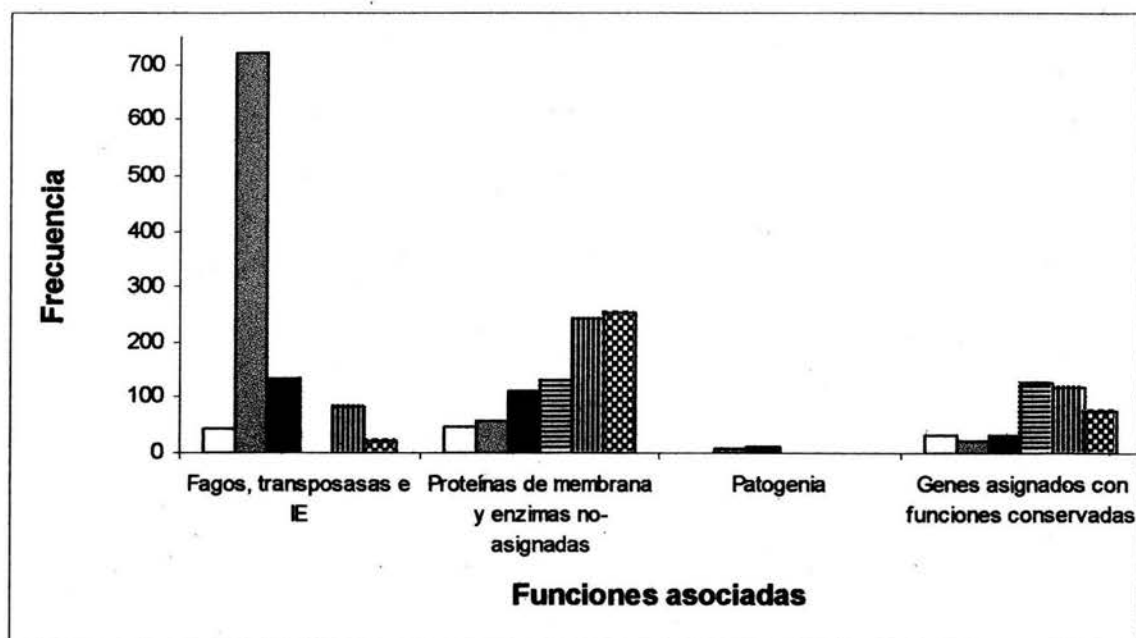


Figura 4. Funciones asociadas a los GTH y a los genes que se perdieron. Las columnas blancas son los datos de los GTH de *E. coli* K12 MG1655, las columnas grises son los datos de los GTH de *E. coli* O157 EDL933 y las columnas negras son los datos de GTH de *S. typhimurium* LT2. Las columnas con líneas horizontales son los datos de los genes que no fueron detectados por el MM en *E. coli* K12 MG1655, las columnas con líneas verticales son los datos de los genes que no fueron detectados por el MM en *E. coli* O157 EDL933. Las columnas blancas con negro son los datos de los genes que no fueron detectados por el MM en *S. typhimurium* LT2. Nótese el número de genes verdaderamente extraños detectados por el modelo (genes de profago). Los genes de profago no detectados por el modelo son probablemente falsos negativos. La presencia de GTH con funciones conservadas podría tratarse de falsos positivos.

Duplicación de Fragmentos de DNA

"Cuando queráis engañar al mundo, decidle la verdad".
Bismarck.

Aproximadamente el 44% de los genes de los tres genomas de enterobacterias analizados tiene por lo menos un gen parálogo (Tabla 11), detectable a partir de los análisis de estructura primaria de sus proteínas (ver Métodos). Este dato indica que la duplicación de material genético está fuertemente seleccionada. Es por ello que se ha postulado como el principal mecanismo bajo el cual las rutas metabólicas se han estructurado a lo largo del tiempo (Bernard y Riley 1995; Lazcano y Miller 1999; Brosius 2003; Hooper y Berg 2003a).

La gran mayoría de los genes duplicados se encuentran conservados en los tres organismos, excepto en *E. coli* O157, en donde se aprecia un aumento considerable en el número de genes duplicados recientemente. En algunos casos podríamos estar hablando de duplicaciones que se dieron hace varias decenas de millones de años, antes de la divergencia de las proteobacterias y (en total 867 parejas).

Organismo	Núm. de genes del organismo que codifican para una proteína	Núm. de genes con al menos una duplicación	Núm. de genes con al menos una duplicación reciente (100MA)
<i>E. coli</i> K 12	4287	1741 (40.6 %)	204 (4.7 %)
<i>E. coli</i> O157	5348	2441 (45.6 %)	702 (13.1 %)
<i>S. typhimurium</i>	4553	2201 (48.3 %)	234 (5.1 %)

Núm. de genes con al menos una duplicación que se comparten en las tres Bacterias	Núm. de genes con al menos una duplicación que se comparten en las <i>E. coli</i>
1288 (que equivalen al 74% de los genes parálogos en K, al 56% en O y al 70.1% en S)	1595 (que equivalen al 91.6% de los genes parálogos en K y al 66.9% en O)

Tabla 11. Datos generales de las duplicaciones en las enterobacterias analizadas. O es *E. coli* O157, K es *E. coli* K 12 y S es *S. typhimurium*.

La distribución de los genes duplicados en las diferentes categorías metabólicas fue distinta a la esperada al azar ($p < 0.0001$). Se realizó un primer análisis global cuantificando todos los posibles genes duplicados y se clasificaron según su ruta metabólica. Es más frecuente encontrar duplicaciones en genes relacionados con transportadores de membrana, transducción de señales, metabolitos secundarios, degradación de xenobióticos y metabolismo de carbohidratos (Tabla 12). Absolutamente todas las vías metabólicas presentaron duplicaciones por arriba del 30% de sus genes (Tabla 12). Este dato es interesante resaltarlo ya que corrobora que la duplicación de genes ha jugado un papel importante en la estructuración de todas las vías metabólicas (por lo menos en un cierto grado que todavía es detectable).

¿En qué vías, entonces, se presentan las duplicaciones recientes? ¿En qué vías aparecen las

duplicaciones más ancestrales que todavía se pueden detectar? Para responder a estas interrogantes se decidió analizar la duplicación de genes desde un punto de vista temporal: a partir de la divergencia entre las *E.coli* y la *Salmonella* y desde la divergencia de todas la proteobacterias γ (Gupta 2000).

Los resultados de este estudio (Tablas 13 y 14) muestran un pequeño número de genes duplicados en *E.coli* K12 y en *S.typhimurium* (204 y 234 respectivamente, equivalente al $\sim 4.9\%$ de todo el genoma en ambos casos); mientras que en la *E.coli* O157 se duplicaron 702 genes (el 13.1% de todo el genoma).

En las tres bacterias los genes que presentaron mayor frecuencia de duplicación reciente fueron los no-asignados (Tabla 13). Esta categoría no figuraba de manera sobresaliente en el análisis global, mas es la primera en importancia en el análisis de duplicaciones nuevas (Tablas 12 y 13). La diferencia entre el $\sim 4.7\%$ de genes duplicados en *E.coli* K12 y *S. typhimurium* con respecto al 13.1% de *E.coli* O157 está dada por un aumento en el número de genes duplicados, precisamente en el grupo de genes no-asignados.

Las otras rutas que retuvieron a los nuevos genes duplicados, en las tres bacterias, exceptuando el caso de los genes no-asignados, son aquellas que según el análisis global presentan las más altas frecuencias de este tipo de genes: transporte a través de la membrana, transducción de señales, metabolismo de carbohidratos (vías periféricas) y biodegradación de xenobióticos (Tablas 12 y 13). Esto podría querer decir que son vías que permiten un gran número de cambios, que aún con el paso del tiempo no se han terminado de estabilizar y que probablemente nunca lo hagan (debido a la aparición de nuevos xenobióticos, sustratos, etc.). Son funciones que se mantienen un alto dinamismo.

Ruta	<i>E.coli</i> K12		<i>E.coli</i> O157		<i>S.typhimurium</i>	
	Frecuencia de genes duplicados	% en esa ruta	Frecuencia de genes duplicados	% en esa ruta	Frecuencia de genes duplicados	% en esa ruta
M. de carbohidratos	159	57.29	154	56.21	201	67.68
M. energía	92	53.18	91	52.17	118	61.46
M. lípidos	25	50.00	52	52.08	29	59.18
M. nucleótidos	36	30.51	37	31.35	45	39.48
M. de otros aminoácidos	36	49.67	35	51.48	40	57.14
M. de carbohidratos complejos	35	33.98	34	34.70	43	43.00
M. de lípidos complejos	31	49.21	30	49.19	29	44.62
M. de cofactores y vitaminas	48	27.75	58	32.58	71	37.77
Metabolitos secundarios	15	60.00	12	57.00	16	72.73
Biodegradación de xenobióticos	33	56.90	35	62.50	31	58.49
Transcripción	44	57.15	40	57.93	46	68.66
Traducción	18	13.84	19	14.50	24	18.46
Clasificación y	12	21.42	24	39.35	47	56.63

<i>Ruta</i>	<i>E.coli K12</i>		<i>E.coli O157</i>		<i>S.typhimurium</i>	
degradación						
Replicación y reparación	14	15.05	16	17.20	22	23.92
Transporte de membrana	337	76.08	338	74.94	324	83.29
Transducción de señales	80	84.21	69	71.03	55	62.50
Movilidad celular	18	30.50	19	34.54	24	30.67
No-asignados	903	33.88	1596	42.91	1247	43.00

Tabla 12. Distribución general de los genes duplicados en las distintas categorías metabólicas en las tres enterobacterias analizadas. Se sombreadon los datos con mayor frecuencia y porcentaje relativo de genes duplicados.

A partir de la divergencia entre *E.coli* y *S.typhimurium*

<i>Ruta</i>	<i>E.coli K12</i>		<i>E.coli O157</i>		<i>S.typhimurium</i>	
	Frecuencia de genes duplicados	% en esa ruta	Frecuencia de genes duplicados	% en esa ruta	Frecuencia de genes duplicados	% en esa ruta
M. de carbohidratos	9	3.24	13	4.75	24	8.08
M. energía	4	2.31	4	2.29	5	2.60
M. lípidos	4	8.00	4	8.33	0	0.00
M. nucleótidos	2	1.70	2	1.69	0	0.00
M. aminoácidos	9	3.62	7	2.86	14	5.36
M. de otros aminoácidos	2	2.76	3	4.41	2	2.86
M. de carbohidratos complejos	1	0.97	1	1.02	1	1.00
M. de lípidos complejos	1	1.59	3	4.92	3	4.62
M. de cofactores y vitaminas	1	0.58	2	1.12	5	2.66
Metabolitos secundarios	0	0.00	0	0.00	0	0.00
Biodegradación de xenobióticos	2	3.45	1	1.79	0	0.00
Transcripción	4	5.20	0	0.00	0	0.00
Traducción	0	0.00	2	1.53	0	0.00
Clasificación y degradación	1	1.79	0	0.00	0	0.00
Replicación y reparación	0	0.00	0	0.00	2	2.17
Transporte de membrana	23	5.19	23	5.10	11	2.83
Transducción de señales	2	2.11	3	3.09	1	1.14
Movilidad celular	0	0.00	0	0.00	2	2.56
No-asignados	160	6.00	658	17.69	187	6.45

Tabla 13. Distribución de los genes duplicados a partir de la divergencia entre *E.coli* y *S.typhimurium*, en las distintas categorías metabólicas en las tres enterobacterias. Se sombreadon los datos con mayor frecuencia y porcentaje relativo de genes duplicados.

Los genes duplicados antes de la especiación entre las proteobacterias γ pertenecen a las vías más ancestrales y conservadas en las bacterias. La tabla 14 muestra el porcentaje de genes que se duplicaron antes de la divergencia de las proteobacterias γ , con respecto al número total de genes con duplicación que aparecen en cada una de las clases. Nótese como en las vías más antiguas (aquellas que probablemente ya estaban presentes en el último ancestro en común de todos los seres vivos) estas duplicaciones explican la mayoría de las duplicaciones detectadas. Resalta el caso de la categoría de traducción, en donde el 100% de las duplicaciones que se encuentran sucedieron antes de la divergencia de las proteobacterias γ . Lo mismo sucede en replicación y reparación con el 78%, movilidad celular con el 61%, metabolismo de cofactores y vitaminas con el 62%, metabolitos secundarios con el 66%, etc. (Tabla 14). Los resultados indican que la duplicación representó un papel importante en estas funciones celulares, pero que aconteció en tiempos muy remotos, previo a la divergencia del último ancestro en común.

<i>Ruta</i>	<i>Total de genes con al menos una duplicación</i>	<i>Total de genes con al menos una duplicación previa a la divergencia de las proteobacterias γ</i>	<i>Porcentaje de los genes duplicados antiguos con respecto al total de genes</i>
M. de carbohidratos	159	46	28
M. energía	92	35	38
M. lípidos	25	11	44
M. nucleótidos	36	14	38
M. aminoácidos	107	58	54
M. de otros aminoácidos	36	19	52
M. de carbohidratos complejos	35	19	54
M. de lípidos complejos	31	9	29
M. de cofactores y vitaminas	48	30	62
Metabolitos secundarios	15	10	66
Biodegradación de xenobióticos	33	12	36
Transcripción	44	20	45
Traducción	18	18	100
Clasificación y degradación	12	7	58
Replicación y reparación	14	11	78
Transporte de membrana	337	135	40
Transducción de señales	80	36	45
Movilidad celular	18	11	61

<i>Ruta</i>	<i>Total de genes con al menos una duplicación</i>	<i>Total de genes con al menos una duplicación previa a la divergencia de las proteobacterias γ</i>	<i>Porcentaje de los genes duplicados antiguos con respecto al total de genes</i>
No-asignados	903	210	23

Tabla 14. Genes duplicados antes de la divergencia de las proteobacterias γ . Se muestra también el porcentaje que abarcan dichos genes con respecto al total de genes duplicados por cada ruta o función celular.

Se hizo un análisis de las posiciones que presentan las parejas de parálogos dentro del mapa metabólico. Los resultados muestran que ambos genes duplicados aparecen generalmente en la misma vía metabólica, en cualquiera que esta sea (en metabolismo central o en rutas accesorias). Son pocos los casos en donde ocurre lo contrario (Tabla 15). Cuando la pareja de genes está presente en la misma vía, tienden también a permanecer en el mismo paso metabólico. Es raro encontrar que dichos genes aparezcan en pasos alternos, y son aun menos los que aparecen en pasos consecutivos

Al realizar un análisis histórico, las parejas de genes que aparecen en funciones celulares que no han aceptado nuevas duplicaciones se localizan en la misma vía (602 casos previos a la divergencia de las proteobacterias γ ; 69%).

Algunas de las parejas más antiguas presentan diversificación de funciones y aparecen en categorías distintas (119 casos antes de la divergencia de las proteobacterias γ , además de 146 casos mal asignados). Hay duplicaciones más recientes que también tienen una diversificación en sus funciones dentro del metabolismo celular (188 casos posteriores a la divergencia de las proteobacterias γ , además de 560 casos mal asignados).

Organismo	Núm. Parejas que están en la misma vía	Núm. Parejas que están en distintas vías	Núm. Parejas que están MAL asignadas
<i>E.coli K 12</i>	2713	301	706
<i>E.coli O157</i>	4068	364	865
<i>S.typhimurium</i>	3902	383	1631

Tabla 15. Distribución funcional de las parejas de genes duplicados. Aparece el número de parejas donde ambos genes están en la misma vía, donde cada uno de ellos está en vías distintas o donde uno sí está asignado y el otro no (mal asignados).

Pérdida de Material Genético

*"La ciencia nos ha prometido la verdad, pero nunca nos ha prometido ni la paz ni la felicidad".
Gustavo Lebon.*

En el tiempo que ha pasado desde que *S.typhimurium* y *E.coli* divergieron, se ha perdido alrededor del 7% de los genes de sus genomas (Tabla 16). La cepa no patógena de *E.coli* es la que ha perdido más genes, por eventos que ocurrieron como parte del proceso de diferenciación entre las propias cepas (Tabla 16).

Un análisis de las características de los genes que se han perdido indica que la gran mayoría de ellos codificaban para enzimas o proteínas de membrana, o son genes de profago crípticos no-funcionales (Tabla 17). Nótese que son también las proteínas que más se transfieren entre bacterias (Figura 4).

Organismo	Núm. Genes del organismo que codifican para una proteína	Núm. Pérdidas
<i>E.coli</i> K 12	4287	385 (8.9 %)
<i>E.coli</i> O157	5348	349 (6.5 %)
<i>S.typhimurium</i>	4553	356 (7.8 %)

Núm. de genes que se perdieron en las tres Bacterias	Núm. de genes que se perdieron sólo en las <i>E.coli</i>
0	267 (que equivalen al 69.3% de los genes perdidos en K y al 76.5% en O)

Tabla 16. Datos generales de los genes perdidos en las enterobacterias analizadas.

Función asociada al gen	Frecuencia de Genes Perdidos			
	<i>E.coli</i>	<i>E.coli</i> K12	<i>E.coli</i> O157	<i>S.typhimurium</i>
Enzima	74	14	15	61
Membrana	73	15	13	44
DNA polimerasa subunidad epsilon putativa	1	0	0	0
Citocromo oxidasa putativa	0	0	0	1
Reguladores transcripcionales	16	2	5	51
Endonucleasas putativas	3	0	0	0
Adenina glicosilasa	0	1	0	0
Fago	21	14	0	22
Chaperona	1	0	0	2
DNA polimerasa II	1	0	0	0
Intergénico	0	0	27	0

Exonucleasas	2	2	0	2
Helicasa putativa	2	6	0	7
Choque térmico	0	0	0	1

Tabla 17. Funciones asociadas a los genes que se perdieron en las enterobacterias analizadas. Se muestran los genes perdidos por cada una de ellas así como los genes perdidos compartidos entre las dos *E. coli*.

Es claro el hecho de que las vías con mayor número de genes perdidos coinciden con aquellas que más se han duplicado recientemente o con aquellas que han tenido mayor llegada de genes por transferencia horizontal (Tablas 10, 12, 13, 18 y 19). Dentro de un genoma, queda claro que existen funciones celulares evolutivamente estables, como el metabolismo central, la traducción, etc. que no aceptan cambio alguno. Algunas veces estas vías pueden perder alguna proteína sin que se vea afectada la adecuación del organismo, pero esto sucede sólo de manera incidental: por ejemplo, las *E. coli* perdieron un fragmento de DNA que contenía varios genes del metabolismo de vitaminas y cofactores que teóricamente no se deberían perder (Tabla 18).

Ruta	<i>E. coli</i> K12		<i>E. coli</i> O157		<i>S. typhimurium</i>	
	Frecuencia de genes perdidos	% en esa ruta	Frecuencia de genes perdidos	% en esa ruta	Frecuencia de genes perdidos	% en esa ruta
M. de carbohidratos	24	8.66	28	10.22	22	7.41
M. energía	7	4.00	6	3.45	4	2.00
M. lípidos	4	8.00	5	10.42	5	10.20
M. nucleótidos	4	3.39	3	2.54	3	2.63
M. aminoácidos	10	4.00	15	6.12	12	4.60
M. de otros aminoácidos	5	6.85	8	11.76	6	8.57
M. de carbohidratos complejos	4	3.88	3	3.00	4	4.00
M. de lípidos complejos	7	11.11	7	11.48	1	1.54
M. de cofactores y vitaminas	18	10.40	16	8.99	3	1.60
Metabolitos secundarios	1	4.00	2	9.52	2	9.00
Biodegradación de xenobióticos	3	5.17	6	10.71	6	11.32
Transcripción	0	0.00	3	4.35	3	4.48
Traducción	1	0.77	1	0.76	0	0.00
Clasificación y degradación	0	0.00	5	8.20	5	6.00
Replicación y reparación	4	4.30	2	2.15	2	2.17
Transporte de membrana	34	7.67	30	6.65	21	5.40

<i>Ruta</i>	<i>E.coli K12</i>		<i>E.coli O157</i>		<i>S.typhimurium</i>	
Transducción de señales	1	1.00	2	2.00	1	1.14
Movilidad celular	1	1.69	1	1.82	0	0.00
No-asignados	402	15.00	367	9.87	188	6.49

Tabla 18. Distribución general de los genes perdidos en las distintas categorías metabólicas en las tres enterobacterias a partir de la divergencia entre los linajes. Se sombreadon los datos con mayor frecuencia y porcentaje relativo de genes perdidos.

<i>Ruta</i>	<i>E.coli K12</i>		<i>E.coli O157</i>	
	Frecuencia de genes perdidos	% en esa ruta	Frecuencia de genes perdidos	% en esa ruta
M. de carbohidratos	6	2,19	10	3,61
M. energía	1	0,57	0	0
M. lípidos	1	2,08	2	4
M. nucleótidos	1	0,85	0	0
M. aminoácidos	1	0,41	6	2,41
M. de otros aminoácidos	1	1,47	4	5,48
M. de carbohidratos complejos	2	2,04	1	0,97
M. de lípidos complejos	0	0	0	0
M. de cofactores y vitaminas	2	1,12	0	0
Metabolitos secundarios	0	0	1	4
Biodegradación de xenobióticos	0	0	3	5,17
Transcripción	0	0	3	3,9
Traducción	0	0	0	0
Clasificación y degradación	0	0	5	8,93
Replicación y reparación	2	2,15	0	0
Transporte de membrana	13	2,88	9	2,03
Transducción de señales	0	0	1	1,05
Movilidad celular	0	0	0	0
No-asignados	102	2,74	67	2,51

Tabla 19. Distribución general de los genes perdidos en las distintas categorías metabólicas a partir de la divergencia entre las *E.coli*. Se sombreadon los datos con mayor frecuencia y porcentaje relativo de genes perdidos.

Como regla general se expone que son los genes accesorios o sin importancia aparente los que se pierden. Estos genes accesorios pertenecen a funciones celulares que cambian mucho y que tienden a aumentar mucho el número de genes (con la misma facilidad se pierden como llegan o se duplican), todo depende de los cambios en las condiciones ambientales y por lo tanto de los requerimientos celulares y las presiones de selección que imperan.

Desentrañando la Historia de los Genes sin Pasado Evidente y como Aprendí a Extraviar la Información

"El conocimiento es un proceso de acumulación de hechos; la sabiduría consiste en su simplificación".
Martín H. Fischer.

En cada genoma analizado, aproximadamente un 40% de los genes no tienen ningún rasgo que nos pudiera indicar si su origen fue por transferencia horizontal, duplicación de material genético, secuencias simples o algún otro mecanismo desconocido. Se trata de un número considerable de genes y por lo tanto de una gigantesca cantidad de información biológica a la cual no tenemos acceso. La mayoría de ellos pertenecen a las funciones celulares más conservadas. Una pregunta crucial se desprende de todo esto: ¿Cómo podríamos desentrañar la historia de estos genes sin pasado evidente?

Richard Dawkins en "the blind watchmaker" (Dawkins 1996), hizo un pequeño experimento de cambios puntuales a la pequeña frase del Hamlet de Shakespeare "METHINKS IT IS LIKE A WEASEL". La probabilidad de generar la frase al azar, tomando los 28 caracteres que conforman la lengua inglesa, sería de 2.72×10^{-47} . Ahora bien, si a cada intento de escribir la frase se le añade un sistema de selección que permita conservar aquellas letras que aparecieron correctamente en la posición, entonces sólo se necesitarían alrededor de 30 ciclos; ésta es la magia de la selección natural. Una idea similar la realizaron experimentalmente Hayashi y sus colaboradores. Ellos generaron una biblioteca de proteínas al azar de tamaño definido y fueron escogiendo aquella proteína que tornaba más patógeno a un profago. Al final de 50 ciclos lograron obtener una secuencia muy similar a la proteína nativa del virus (Hayashi *et al.* 2003).

La idea de Dawkins podría ser reinterpretarla de acuerdo al interés de desentrañar el pasado de las moléculas. La lógica a seguir es la siguiente: Existen modelos bien definidos de detección de secuencias simples, duplicación de material genético, y ahora gracias a una de las aportaciones de este trabajo, de genes de transferencia horizontal reciente. Cada modelo está definido bajo ciertos índices y parámetros que nos enmarcan lo que entendemos por cada uno de estos fenómenos. El procedimiento es generar modelos probabilísticos para cada tipo de rearrreglo (secuencias simples, duplicación de genes y transferencia horizontal) y determinar qué modelo explica mejor la situación de un gen cuya historia nos es desconocida. Para determinar que modelo es el adecuado, se escoge un gen al azar y se procede a hacerle mutaciones puntuales seleccionadas, de tal manera que vaya cambiando hasta volverse una secuencia simple, hasta que encuentre un homólogo en su propio genoma, o hasta que posea un arreglo de bases lo suficientemente distinto como para que se pudiera identificar como un gen de transferencia horizontal.

Los resultados son concluyentes: el modelo que explica mejor a los genes sin historia conocida es, invariablemente, la transferencia horizontal. Pero existen varios inconvenientes insalvables a la metodología utilizada que pudieran estar afectando los resultados. El primero es que las mutaciones se realizaron en cualquier parte del gen sin importar si la función de éste se veía afectada. Segundo, el modelo de transferencia horizontal tiene una gran ventaja y es que de acuerdo a como están planteados los modelos

probabilísticos, siempre se necesitarán muchos menos cambios para que un gen ya no se parezca a su genoma que para que adquiriera otra condición (un parálogo). De esta forma se aprecia que, aun si poseemos modelos robustos que nos identifican cierto tipo de rearrreglos genómicos, éstos no son suficientes para jugar con las probabilidades de aparición de los distintos fenómenos.

Organismo donador del Gen	Media	DS
<i>A.pernix</i>	6.41	1.78
<i>S.solfataricus</i>	9.71	3.36
<i>M.jannaschii</i>	13.47	1.04
<i>D.radiodurans</i>	5.58	2.28
<i>C.pneumoniae</i>	3.52	1.74
<i>M.tuberculosis</i>	2.73	3.18
<i>M.genitalium</i>	9.95	3.64
<i>C.actetobutylicum</i>	9.26	1.18
<i>S.aureus N315</i>	7.33	2.99
<i>B.subtilis</i>	1.06	1.36
<i>C.crescentus</i>	3.86	1.00
<i>M.loti</i>	2.35	2.42
<i>V.cholerae</i>	1.61	3.1
<i>Y.pestis</i>	1.59	2.59
<i>H.influenzae</i>	2.12	1.76

Tabla 20. Porcentaje de cambios necesario para que un gen de estos genomas se mimetice con el entorno genómico de *E.coli*.

Aunque la metodología de este experimento falló en su propósito central, sirvió para determinar el número de cambios que necesitaría acumular un gen de transferencia horizontal recién llegado para parecerse más a su nuevo genoma y por lo tanto volverse indetectable (extraviarse ante el análisis del modelo de Markov). Se hicieron dos experimentos distintos. El primero consistió en mutar 20 genes tomados al azar de 15 genomas, con el fin de calcular el número de cambios necesarios para que se parecieran al contexto genómico de *E.coli* K12 (Tabla 20). En el segundo experimento se generaron mutaciones puntuales a dos distintos tipos de genes: aquellos pertenecientes a vías metabólicas que no han aceptado GTH recientemente en las enterobacterias y aquellos pertenecientes a vías en donde han aparecido notoriamente. Los genes se tomaron al azar de 9 genomas y se mutaron hasta que se lograron mimetizar en el entorno del genoma de *E.coli* K12 (Tabla 21). Esto último se hizo con el fin de definir si el modelo de Markov encontraba con mayor facilidad ciertos tipos de genes porque eran éstos los que más difícilmente se mimetizaban con su entorno.

Organismo Donador del Gen	Genes de Rutas sin TH		Genes de Rutas con TH	
	Media	Ds	Media	Ds
<i>A.aeolicus</i>	2.35	1.52	3.56	1.57
<i>A.pernix</i>	4.79	1.56	3.58	0.71
<i>D.radiodurans</i>	2.88	1.88	1.90	1.42
<i>S.aureus n315</i>	5.04	2.03	6.11	2.21
<i>B.subtilis</i>	0.44	0.86	0.09	0.02
<i>C.crescentus</i>	3.44	1.25	3.06	2.01
<i>C.jejuni</i>	5.76	2.85	8.69	1.49
<i>Y.pestis</i>	0.17	0.44	0.21	0.21
<i>H.influenzae</i>	1.98	0.44	2.44	1.18

Tabla 22. Porcentaje de cambio que requirieron acumular genes, provenientes de los organismos que se enlistan, pertenecientes a rutas metabólicas que no tuvieron indicios de GTH en *E.coli* y rutas que sí los tuvieron para mimetizarse con el entorno genómico de *E.coli*.

Los resultados muestran que la cercanía filogenético entre el organismo donador y el organismo aceptor del gen determinará el número de mutaciones que debe acumular el gen para que se vuelva indetectable al modelo de Markov (Tablas 21 y 22). En organismos muy lejanos, el total de modificaciones pueden llegar a ser del 20%. En organismos más cercanos o en los *Bacillus*, el total de modificaciones es mínimo. Es importante subrayar lo siguiente: los genes de un mismo genoma necesitan un número de cambios similar para parecerse a genes de otro genoma sin importar que función realicen en la célula. Nuevamente hay un problema con el planteamiento de la metodología y es que los cambios se realizaron en todo el gen por igual sin que importase la pérdida de la función o la estructura.

CONCLUSIONES

"La ignorancia es preferible al error, y se halla menos lejos de la verdad el que no cree nada que el que cree algo falso".
Thomas Jefferson.

Sobre las Secuencias Simples

La presencia de este tipo de secuencias está dada en dos clases principales de proteínas: asociadas a membrana (primordialmente) y enzimas citosólicas. Las SeSp son una clara amplificación de los aminoácidos más comunes del proteoma de *E.coli*, seguramente porque su formación está determinada por el barrido de la polimerasa (Bzymek & Lovett 2001) y en este proceso se insertan los aminoácidos de acuerdo a sus abundancias relativas. Las SeSp se encuentran casi por igual en alfas hélices como en asas. Su ausencia en láminas beta es un asunto no resuelto. En las proteínas de membrana no siempre aparecen formando parte de los dominios intermembranales, sino que en muchos casos se encuentran en los dominios externos, lo que podría indicar una posible participación en la función de las proteínas. No hay ningún trabajo que ayude a explicar la real importancia de las SeSp en los genomas. Sólo Bayliss, Dixon y Moxon (Bayliss *et al.* 2004) se habían aventurado a decir que las SeSp permitían el desarrollo y diversificación de la patogenicidad en las bacterias. Los resultados de este trabajo muestran que no hay relación alguna, por lo menos en las enterobacterias, de las SeSp y sus proteínas de patogenicidad (Dobrindt y Reidl 2000). ¿Cuál es la función que desempeñan las SeSp que permite su mantenimiento? La pregunta se mantiene sin respuesta clara.

No hay selección *de novo* de SeSp: las diferencias entre las distribuciones de genes con SeSp en las enterobacterias se deben a la pérdida de ortólogos con SeSp o a la ganancia de genes por duplicación o transferencia horizontal que ya tenían una SeSp. ¿Entonces cuándo se formaron las más de trescientas SeSp que se detectan? Muchas de ellas llegaron por transferencia horizontal reciente (algunas proteínas de profagos las presentan). Otras se conservan (aunque sea sólo el hecho de tener una SeSp) desde antes de la divergencia del último ancestro en común ya sea a partir de diversificación de las proteobacterias y o incluso antes. Ejemplos muy interesantes son la subunidad c de la ATP sintetasa o la proteína L1 del ribosoma; las SeSp están presentes en casi todos los linajes bacterianos. Estas proteína en particular tienen, además, una cobertura de SeSp superior al 20%. La antigüedad de estas dos proteínas sugiere que algunas de las SeSp que aún se detectan en la actualidad podrían ser las huellas de un mundo remoto donde las proteínas eran más sencillas que ahora. Un último dato que es importante resaltar es la presencia de una proteína de membrana (STM4222 que presentó SeSp ricas en serinas en un 56% de su secuencia) y que podría ser evidencia de que se pueden obtener proteínas con funciones muy complejas formadas meramente de secuencias simples. Más experimentos son requeridos para resolver los problemas pendientes.

Sobre la Transferencia Horizontal

Las metodologías probadas en este trabajo han sido ampliamente usadas en los últimos años (Aravind et al. 1998; Hayes & Borodovsky 1998; Kyrpides et al. 1999; Ochman & Lawrence 1998; Garcia-Vallvé et al. 2000; Hooper & Berg 2002; Snel et al. 2002; Daubin et al. 2003b) para detectar genes transferidos horizontalmente, pero han sido empleadas sin un control experimental que pudiera ayudar a determinar si realmente estaban funcionando de una manera correcta. Al usar un control experimental, la adición artificial de genes extraños conocidos al genoma de *E.coli* K12 MG1655 y *E.coli* O157 EDL933, la capacidad de detección de los modelos fue muy baja. Sólo el modelo de Markov construido fue robusto y consistente en las detecciones, fallando, sin embargo, cuando los genes pertenecían a organismos muy cercanos.

La capacidad de detección de los modelos fue aún más baja cuando se utilizó todo el genoma para hacer los cálculos de los parámetros (media y desviación estándar), en lugar de usar sólo a los genes conservados. Estos resultados se explican si consideramos que la mezcla de genes conservados y no-conservados (donde muchos genes horizontales estarían presentes) produce una media genómica errónea y una desviación estándar más grande. Esto se apoya con los resultados obtenidos en la cepa patógena de *E.coli*: su genoma reveló la mayor cantidad de GTH, y es en este organismo donde las tasas de detección decrecieron marcadamente cuando el genoma completo fue utilizado (Tabla 8).

Se han detectado genes a través de una aproximación composicional con un umbral de cuatro desviaciones estándar que falla en detectar genes provenientes de organismos cercanos (Tabla 9). Esto quiere decir que los genes detectados en las tres enterobacterias analizadas pertenecieron a organismos muy distantes, y que aún se asemejan a su contexto genómico previo, contrario a lo propuesto (Daubin et al. 2003a). Por lo tanto, las barreras que impiden el intercambio de genes entre bacterias no sólo son especie específicos sino que, por el contrario, también son ambientales.

Se encontró una adquisición independiente considerable de genes en las tres bacterias analizadas, especialmente en la cepa patógena de *E.coli*. Se descubrieron también genes adyacentes entre tres y cuatro desviaciones estándar. Esto podría tener dos explicaciones: algunos de los genes transferidos han acumulado mutaciones tras su llegada o estos genes presentan arreglos de bases cercanos a los encontrados en los genes conservados. Cualquiera que sea la respuesta, los resultados prueban que este complemento metodológico a la primera búsqueda, permitió el descubrimiento de genes transferidos más ocultos.

La aplicación del modelo de Markov podría ser una herramienta poderosa y acertada para resolver una pregunta biológica muy interesante: Si los genes no-compartidos entre dos organismos hermanos se deben a la ganancia por transferencia horizontal o a la pérdida de genes. Esta tesis se apoya en el hecho de que muchos de los GTH son claramente foráneos (genes relacionados con fagos; Figura 4) y, además, hay muchos más genes asignados en el grupo de genes que se perdieron (Figura 4). La exactitud del modelo de Markov se puede corroborar al contar el número de probables falsos positivos y falsos negativos; algunos genes foráneos (genes de profago) que no fueron detectados por el modelo y los genes asignados que fueron detectados como GTH (Figura 4).

La idea de un gran intercambio de genes genera la siguiente pregunta, ¿todos los genes están igualmente sujetos a ser transferidos? (Ochman & Lawrence 1997). Si esta hipótesis resultara cierta, el DNA foráneo reemplazaría todos los genes heredados verticalmente de un genoma en pocos millones de años (Ochman & Lawrence 1997). Por ello, se supuso que la historia de la vida no podría representarse correctamente como un árbol (Doolittle 1999). Muchos trabajos han intentado corroborar esta hipótesis (de la Cruz & Davies 2000; Gogarten et al. 2002). Recientemente, se comprobó que los GTH detectados a través de un método PD codificaban solamente para proteínas no-asignadas (Daubin et al. 2003b). Este dato muestra que aquellos genes más esenciales en un genoma, no están sujetos a este fenómeno. Los datos de este trabajo corroboran la naturaleza no-asignada de los GTH. Sin embargo, un análisis más refinado muestra que podrían estar involucrados en la interacción directa del organismo y su ambiente: proteínas de membrana, enzimas, proteínas de choque térmico y ácido y proteínas relacionadas con la patogenicidad. Esta distribución funcional es constante en las tres enterobacterias analizadas. Por lo tanto, estos genes podrían estar involucrados en mantener la adecuación de las bacterias. Este aspecto ha sido ampliamente observado en las poblaciones de bacteria que habitan en ambientes contaminados por xenobióticos (revisado por Top & Springael 2003). Más experimentos son requeridos para develar este problema.

Sobre las Duplicaciones de Genes

La duplicación de genes ha sido sin duda alguna uno de los principales mecanismos para aumentar el tamaño y complejidad de los genomas. Este fenómeno tiene una gran ventaja sobre los demás y es que de inmediato genera una nueva copia de un gen que ya ha sido probado en ese genoma. Aunque los resultados claramente exponen que por lo general los genes que se duplican se quedan en la misma vía metabólica en el mismo paso, este proceso ha sido seleccionado en todas las rutas metabólicas, en todas las funciones celulares, en genes de reciente adquisición (GTH) como en genes sumamente antiguos (genes involucrados en el aparato de traducción, replicación y reparación, movilidad celular). Lo que nos dice que es un proceso que ha ocurrido desde que las primeras células existieran. Las vías más antiguas presentan las duplicaciones ancestrales también, son pocos o nulos los casos de duplicaciones recientes en ellas.

El principal problema con la duplicación de genes es la sobreestimación actual del fenómeno debido que los análisis se hacen atemporales. De tal manera que al analizar un genoma se encuentran genes duplicados de un enorme intervalo de tiempo. Cuando se analizan desde un punto de vista temporal, por ejemplo a partir de la divergencia de *S.typhimurium* y *E.coli*, el número de genes duplicados se reduce notoriamente, estos genes en su mayoría llegaron por un evento de transferencia horizontal reciente.

Los resultados muestran que muchos de los genes que se duplican tienden a permanecer en el mismo paso metabólico sin importar que tan antiguas sean las parejas. De hecho, la mayoría de las parejas que están en vías distintas tienen apariciones más recientes, digamos tras la divergencia de las proteobacterias γ . Estos datos sugieren que en los genes parálogos no se aprecia la adquisición de nuevas funciones conforme el tiempo ha transcurrido. Pero en realidad, un análisis más profundo permite deducir que cuando hay una duplicación de genes existen tres destinos a largo plazo: 1) la permanencia en el mismo

paso metabólico. 2) la pérdida de alguna de las copias. 3) la diversificación hacia otra función a través de un fenómeno tipo patchwork (vista como su aparición en un paso no-consecutivo de la misma vía o en otra vía metabólica distinta). De esta manera, en estos dos últimos casos se pierde en muy poco tiempo el registro de paralogía entre los genes, lo que explica la caída en el número de parejas antiguas con diferentes funciones.

El análisis de los genes duplicados a partir de la divergencia de las enterobacterias analizadas, muestra que en la actualidad existen vías que ya no aceptan nuevas duplicaciones, estas vías tienen que ver con aquellas funciones celulares más ancestrales y establecidas. Es interesante resaltar que aquellas vías que en los últimos milenios han permitido la amplificación de algunos de sus genes, son también aquellas que históricamente han permitido más duplicaciones y que, por lo tanto, representan las vías más plásticas relacionadas con funciones auxiliares (Hooper y Berg 2002). También se aprecia que los genes que más se han duplicado recientemente son los genes de transferencia horizontal (Hooper y Berg 2003). Lo anterior implica que para que un gen duplicado se seleccione en las enterobacterias actuales debe de tener bajas presiones de selección (GTH) o estar involucrado en funciones auxiliares, como aquellas que median la interacción del organismo con su ambiente (genes de transporte a través de membrana y metabolismo de xenobióticos).

Sobre la Pérdida de Material Genético y la Reconstrucción de Historias

La forma de plantear la pérdida de genes en este trabajo podría contener falsos positivos. Todos aquellos GTH cuyos arreglos de bases se parecieran al del genoma analizado, fueron irremediablemente considerados como pérdidas en los otros linajes. Aún así, dada la capacidad de detección probada del modelo de Markov (Tablas 8 y 9), se puede suponer que la gran mayoría de los genes que aquí se asumen como pérdidas, realmente lo son. De tal manera que los datos son robustos a este respecto.

Los resultados son claros: la pérdida es el segundo evento que ha sido más frecuente (en número de genes involucrados) a partir de la divergencia entre las enterobacterias analizadas. Los genes que más se pierden pertenecen a aquellas vías que más se han duplicado en los últimos tiempos, así como aquellas que más han incorporado GTH. Este dato demuestra que la transferencia horizontal ha sido aún mayor de la que detectamos en la actualidad. Un caso concreto serían los genes de profago, éstos suelen llegar masivamente a algunos organismos, pero también son una de las categorías que más se pierden. La respuesta a estas cuestiones es sencilla: los mismos mecanismos que producen el aumento del material genético por duplicación, pueden también derivar en pérdida de fragmentos; y, puesto que los que más se duplican son los GTH, se concluye que todas aquellas vías que aceptan GTH o que simplemente se duplican más, también tenderán a perderse.

Lo interesante es la enorme cantidad de genes que se han perdido sin que la adecuación de los organismos se viera afectada. Esto se puede explicar si: a) los genes que se pierden carecen de importancia dentro de la célula, así como se duplican o transfieren, también se pueden perder. Un ejemplo serían las proteínas crípticas de profago; b) son genes cuya aportación a la adecuación del organismo es temporal y

depende de la tasa de cambio del medio exterior, es decir, estos genes sólo desempeñan una función importante en determinados circunstancias ambientales (cambios en las variables externas). Pasado ese momento, se vuelven prescindibles porque otros genes han tomado ya su lugar. Unos ejemplos serían algunas las proteínas de membrana, las proteínas relacionadas con patogenicidad y enzimas de rutas nuevas como es el caso del metabolismo de xenobióticos. Más investigaciones teóricas y experimentales son requeridas para resolver este asunto.

Al respecto del ensayo de reconstrucción de historias de genes sin un pasado aparente, es claro que los modelos que se han diseñado aún no tienen la capacidad suficiente para estudiar correctamente este tipo de información. La idea de competencia entre los modelos no es mala, pero en muchos casos la información se encuentra totalmente perdida y eso imposibilita que los modelos puedan desarrollarse correctamente. Lo anterior implica que, en este tipo de ejercicios teóricos, siempre gane el modelo que tiene un camino más seguro, en este caso sería la transferencia horizontal (vista como un rearrreglo de bases). Este modelaje evolutivo permitió calcular el número de mutaciones promedio que requiere un GTH para mimetizarse con su nuevo entorno. Lo interesante es que el porcentaje de cambio es constante para todos los genes de un determinado genoma. Este dato fortalece la idea de que hay una cierta tendencia a que los genes de un genoma se parezcan en índices de Markov, lo que ratifica que el modelo de Markov construido es la aproximación correcta y que, además, podría ser aplicable en cualquier caso.

Sobre la Evolución de las Rutas Metabólicas y los Procariontes que las Poseen

Tras millones de años de vida en la Tierra hay varias vías metabólicas que ya no aceptan ninguna alteración en su estructura. Todas ellas están universalmente compartidas puesto que todos los organismos que las presentan son descendientes de aquellas poblaciones exitosas que las reunieron y que desplazaron a sus compañeras de ambiente. ¿Cuáles han sido las vías que no han dejado de aceptar cambios? La respuesta es clara a partir de los datos encontrados: aquellas vías cuyos genes codifican para proteínas neutras, es decir, su presencia es prescindible (ej. proteínas de profago); o las vías cuyas proteínas presentan funciones que median la interacción del organismo con su ambiente (proteínas de membrana, proteínas de choque, enzimas de las rutas de xenobióticos, etc.).

Cabe señalar que algunas veces pueden llegar a transferirse o perderse genes de las funciones más esenciales sin que la adecuación de los organismos decaiga. Estos genes son la clara muestra de que todos los genes están sujetos con la misma probabilidad a la pérdida o a la transferencia. Claro que esto no significa que igualmente se seleccionen.

Se puede decir que los procariontes actuales presentan dos clases de genes muy distintos: a) genes con funciones esenciales, cuyas duplicaciones, transferencias o pérdidas afectan irremediablemente al organismo y b) genes con funciones accesorias que están en constante cambio con respecto al medio exterior. Este grupo de genes permiten que las bacterias se mantengan adaptadas.

La idea de estas dos clases de genes (los reemplazables y los inamovibles) y la actual dinámica genómica en las poblaciones bacterianas, permiten la elaboración de un argumento final sobre la evolución

general de las rutas metabólicas. Esta idea puede ser mejor comprendida si se toman en cuenta, además de los datos aportados por este trabajo, los obtenidos en las investigaciones realizadas en poblaciones de bacterias expuestas a xenobióticos (Top y Springael 2003) y que recuerdan los planteamientos teóricos propuestos por Woese (Woese 1998, Woese 2002).

Para algunos xenobióticos como atrazina (De Souza et al. 1998) y nitrotoleno (Snellinx et al. 2003), la degradación es a veces llevada a cabo por distintas especies de bacterias, cada una de las cuales cataliza una parte de la vía, a través de una enzima de invención independiente. Sin embargo, los mismos genes se encuentran también, para ambos casos, combinados en una sola bacteria, la cual es capaz de mineralizar ese mismo compuesto por ella misma.

Este proceso se puede extrapolar a la evolución de vías metabólicas de la siguiente manera: algunos individuos de alguna población de bacterias presentan (por duplicación y subsiguiente mutación puntual) un gen novedoso que permite la catálisis de algún nuevo compuesto del medio exterior, desechando algún metabolito secundario. Dicha aparición aumenta la adecuación de ese individuo y su descendencia. El mismo proceso sucede en otras poblaciones de células que desarrollan proteínas que degradan los sucesivos metabolitos secundarios. De esta forma, en una comunidad de células, diferentes organismos presentan una nueva vía metabólica completa compartida.

El transporte horizontal permite que estas nuevas proteínas se ensamblen en una misma vía dentro de una sola célula, la cual presenta ahora toda una nueva vía metabólica. Pero como el transporte horizontal es poco preciso en el tipo de genes que involucra, ya que cualquier región del genoma está sujeta a este fenómeno, es necesario que las tasas de transferencia sean altísimas para que, tras una rigurosa selección natural, las enzimas de la vía converjan en una sola población. Esto es exactamente lo que se aprecia de este trabajo. De esta forma se advierte como la transferencia de genes es un mecanismo ampliamente utilizado por las bacterias para intercambiar genes exitosos. Como las transferencias son muy altas, y hay una limitante en el tamaño del genoma de una bacteria, los mecanismos que ocasionan la pérdida de material genético están desregulados y por lo tanto las tasas de pérdida son igualmente elevadas. Si la presión de selección se mantiene, la adecuación del organismo con la vía completa terminará siendo mayor al conjunto de especies que la comparten, puesto que no depende de nadie para sobrevivir. De esta manera, y tras muchos años, esta población de bacterias desplazará a las otras. En ese momento, la vía se vuelve intransferible, inmutable e imperdible y toda su descendencia la presentará.

Como se podrá apreciar, la duplicación de genes genera el material genético donde se ensayarán las nuevas funciones a través de cambios puntuales. La transferencia horizontal permite que inventos aislados se conjuguen en una nueva vía en una sola población. Este modelo de evolución, además que se ha observado en la naturaleza, es más robusto ya que es poco probable que la estructuración de todas las funciones celulares haya ocurrido siempre en una sola población.

REFERENCIAS

Libros

- Darwin, C. El Origen de las Especies. Ed. Editores Mexicanos Unidos. México. 1995.
 Dawkins, R. The blind watchmaker. Ed. Norton. USA. 1996.
 Oparin, A. El Origen de la Vida. Ed. Editores Mexicanos Unidos. 5ta reimpresión. México. 1992.

Artículos Científicos

- Achaz, G; Rocha, EPC; Netter, P & Coissac. (2002). Origins and fate of repeats in bacteria. *Nucleic Acids Research*. 30:2987-2994.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389-3402.
- Aravind, L; Tatusov, RL, Wolf, YI; Walker, DR; Koonin, EV. (1998) Evidence of massive gene exchange between archaeal and bacterial hyperthermophiles. *TIG*. 14:442-444.
- Bayliss, CD; Dixon, KM; & Moxon, RE. (2004). Simple sequence repeats (microsatellites): mutational mechanisms and contributions to bacterial pathogenesis: A meeting review. *FEMS Immunology and Medical Microbiology*. 40 : 11-19.
- Becerra, A; Cocho, G; Delaye, L & Lazcano A. Simple sequences: It is something you have, whether you like or not. (en prep).
- Berg, OG & Kurland, CG. (2002). Evolution of microbial genomes: sequences acquisition and loss. *Mol Biol Evol*. 19:2265-2276.
- Bernard, L & Riley, M. (1995). Widespread protein sequence similarities: Origins of *Escherichia coli* genes. *Journal of Bacteriology*. 177:1585-1588.
- Blattner, F.R., Plunkett, G,3rd., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. & Shao, Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*. 277, 1453-1474.
- Borwn, J. (2003) Ancient horizontal gene transfer. *Nature genetics reviews*. 4:121-132.
- Brookfield, JFY. Gene duplication: the gradual evolution of functional divergence. *Current Biology*. 13:R229-R230.
- Brosius, J. (2003) Gene duplication and other evolutionary strategies: from RNA world to the future. *Journal of Structural and functional genomics*. 3:1-17.
- Clarke, GDP; Beiko, RG; Ragan, MA & Charlebois, RL. (2002). Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and distance matrix based on mean normalized BLASTP scores. *J Bacteriology*. 184:2072-2080.
- Daubin, V., Lerat, E. & Perrière, G. (2003) The source of laterally transferred genes in bacterial genomes. *Genome Biol*. 4, R57.1-R57.12.
- Daubin, V., Moran, N.A. & Ochman, H. (2003) Phylogenetics and the cohesion of bacterial genomes. *Science*. 301, 829-832.
- Daubin, V. & Perrière, G. (2003) G+C3 structuring along the genome: a common feature in prokaryotes. *Mol. Biol. Evol*. 20, 471-483.
- de la Cruz, F & Davies, J. (2000). Horizontal gene transfer and the origin of species: lessons from bacteria. *TRENDS in Microbiology*. 8: 128-133.
- de Souza, ML; Seffernick, J; Martinez, B; Sadowsky, MJ & Wackett, LP (1998). The atrazine catabolism genes atzABC are widespread and highly conserved. *J Bacteriol*. 180: 1951-1954.
- Dobrindt, U. & Reidl, J. (2000) Pathogenicity islands and phage conversion: evolutionary aspects of bacterial pathogenesis. *Int. J. Med. Microbiol*. 290, 519-527.
- Doolittle, W.F. (1999) Phylogenetic classification and the universal tree. *Science*. 284, 2124-2128.
- Doolittle, WF. (1999). Lateral genomics. *Trends Cell Biol*. 9:M5-8.
- Garcia-Vallvé, S., Romeu, A. & Palau, J. (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res*. 10, 1719-1725.

- Gogarten, JP; Dollittle, WF & Lawrence, JG. (2002). Prokaryotic evolution in light of gene transfer. *Mol Biol Evol.* 19:226-2238.
- Guindon, S. & Perrière, G. (2001) Intra-genomic base content variations is a potential source of biases when searching for horizontally transferred genes. *Mol. Biol. Evol.* 18, 1838-1840.
- Gupta, RS. (2000). The phylogeny of protobacteria: relationships to other eubacterial phyla and eukaryotes. *FEMS Microbiology Reviews.* 24:367-402.
- Gur-Arie, R; Cohen, CJ; Eital, Y; Shelef, L; Hallerman, EM & Kashi, Y. (2000). Simple sequence repeats in *Escherichia coli*: abundance, distribution, composition, and polymorphism. *Genome research.* 10:62-71.
- Hayashi, Y; Sakata, H; Makino, Y; Urabe, I & Yomo, T. (2003). Can an arbitrary sequence evolve towards acquiring a biological function? *J Mol Evol.* 56:162-198.
- Hayes, WS & Borodovsky, M. (1998). How to interpret anonymous bacterial genome: machine learning approach to gene identification. *Genome research.* 8:1154-1171.
- Hooper, SD & Berg, OG. (2002). Gene import or deletion: a study of the different genes in *Escherichia coli* strains K12 and O157:H7. *J Mol Evol.* 55:734-744.
- Hooper, SD & Berg, OG. (2003a). On the nature of gene innovation: duplication patterns in microbial genomes. *Mol Biol Evol.* 20: 945-954.
- Hooper, SD; Berg, OG.(2003b). Duplication is more common among laterally transferred genes than among indigenous genes. *Genome biology.* 4:R48.1-48.9
- Jain, R; Rivera, MC; Moore, JE & Lake, JA. (2003). Horizontal gene transfer accelerates genome innovation and evolution. *MBE.* 20:1598-1602.
- Kondrashov, FA; Rogozin, IB; Wolf, YI & Koonin EV. (2002). Selection in the evolution of gene duplications. *Genome biology.* 3:1-9.
- Koonin, EV; Makarova, KS & Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol.* 55:709-742.
- Koski, L.B., Morton, R.A. & Golding, G.B. (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol. Biol. Evol.* 18, 404-412.
- Krypidis, NC & Olsen, GJ. (1999). Archaeal and bacterial hyperthermophiles horizontal gene exchange or common ancestry? *TIG.* 15:298-299.
- Kunin, V & Ouzounis, CA. (2003). The balance of driving forces during genome evolution in prokaryotes. *Genome Research.* 13:1589-1594.
- Kurland, CG. (2000). Something for everyone. *EMBO reports.* 1:92-95.
- Kyte, J & Doolittle, R.F. (1982). A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.* 157: 105-132.
- Lake, JA & Rivera, MC. (2004). Deriving the genomic tree of life in the presence of horizontal gene transfer: Conditioned reconstruction. *Mol. Biol. Evol.* 21:681-690.
- Lawrence & J.G. and Ochman, H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* 44, 383-397.
- Lawrence, J.G. & Ochman, H. (1998) Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. USA* 95, 9413-9417.
- Lawrence, J.G. & Ochman, H. (2002) Reconciling the many faces of lateral gene transfer. *Trends Microbiol.* 10, 1-4.
- Lazcano, A; Miller, SL. (1999). On the origin of metabolic pathways. *J Mol Evol.* 49:424-431.
- Lynch, M; Conery, JS. (2003). The origins of genome complexity. *Science.* 302:1401-1404.
- McClelland, M., Sanderson, K.E., Spieth, J., Clifton, S.W., Latreille, P., Courtney, L., Porwollik, S., Ali, J., Dante, M., Du, F., Hou, S., Layman, D., Leonard, S., Nguyen, C., Scott, K., Holmes, A., Grewal, N., Mulvaney, E., Ryan, E., Sun, H., Florea, L., Miller, W., Stoneking, T., Nhan, M., Waterston, R. & Wilson, R.K. (2001) Complete genome sequence of *Salmonella enterica* serovar *Typhimurium* LT2. *Nature.* 413, 852-856.
- Moran, NA. (2003). Tracing the evolution of gene loss in obligate bacterial symbionts. *Current opinion in Microbiology.* 6:512-518.
- Myers, E & Miller, W. (1988). Optimal alignments in linear space. *CABIOS* 4. 1:11-17.
- Ohno, S. (1970). Evolution by gene duplication. Springer-Verlag, Heidelberg, Germany.
- Perna, N.T., Plunkett, G.3rd., Burland, V., Mau, B., Glasner, J.D., Rose, D.J., Mayhew, G.F., Evans, P.S., Gregor, J., Kirkpatrick, H.A., Posfai, G., Hackett, J., Klink, S., Boutin, A., Shao, Y., Miller, L.,

- Grotbeck, E.J., Davis, N.W., Lim, A., Dimalanta, E.T., Potamouisis, K.D., Apodaca, J., Anantharaman, T.S., Lin, J., Yen, G., Schwartz, D.C., Welch, R.A. & Blattner, F.R. (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*. 409, 529-533.
- Ragan, M.A. (2001) On surrogate methods for detecting lateral gene transfer. *FEMS Microbiol. Lett.* 10, 1-4.
- Ragan, M.A. (2002) Reconciling the many faces of lateral gene transfer: response. *Trends Microbiol.* 10, 4.
- Ragan, MA & Charlebois, RL. (2002). Distributional profiles of homologous open reading frames among bacterial phyla: implications for vertical and lateral transmission. *International journal of systematic and evolutionary microbiology.* 52:777-787.
- Rison, SCG & Thornton JM. (2002). Pathway evolution, structurally speaking. *Current Opinion in Structural Biology.* 12:374-382.
- Roelofs, J & Van Haastert, JM. (2001). Gene lost during evolution. *Nature.* 411:1013-1014.
- Romero, D & Palacios, R. (1997). Gene amplification and genomic plasticity in prokaryotes. *Annu Rev Genet.* 31:91-111.
- Rost, B & Sander, C. (2000). Third generation prediction of secondary structures. *Methods Mol Biol.* 143:71-95.
- Salzberg, SL; White, O; Peterson, J & Eisen, JA. (2001). Microbial genes in the human genome: lateral transfer or gene loss? *Science.* 292:1903-1906.
- Sharp, P.M. & Li, W.H. (1987) The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research.* 15, 1281-1295.
- Snel, B; Bork, P & Huynen, MA. (2002). Genomes in flux: the evolution of archaeal and protobacterial gene content. *Genome research.* 12:17-25.
- Springael, D & Top, EM. (2004). Horizontal gene transfer and microbial adaptation to xenobiotics: new types of mobile genetic elements and lessons from ecological studies. *TRENDS in Microbiology.* 12:53-58.
- Snellinx, Z; Taghavi, S; Vangronsveld, J & van der Lelie D. (2003). Microbial consortia that degrade 2,4-DNT by interspecies metabolism: isolation and characterization. *Biodegradation.* 14:19-29.
- Stanhope, MJ; Lupas, A; Italia, MJ; Koretke, KK; Volker, C & Brown JR. (2001). Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature.* 411:940-944.
- Top, EM & Springael, D. (2003). The role of mobile genetic elements in bacterial adaptation to xenobiotic organic compounds. *Current Opinion in Biotechnology.* 14:262-269.
- van der Meer, JR & Sentchilo, V. (2003). Genomic islands and the evolution of catabolic pathways in bacteria. *Current Opinion in Biotechnology.* 14:248-254.
- Wang, B. (2001) Limitations of compositional approach to identifying horizontally transferred genes. *J. Mol. Evol.* 53, 244-250.
- Woese, C. (1998) The universal ancestor. *PNAS.* 95:6854-6859.
- Woese, C. (2002) On the evolution of cells. *PNAS.* 99:8742-8747.
- Wootton, J. (1994). Non-globular domains in protein sequences: automated segmentation using complexity measures. *Computers Chemistry.* 18:269-285.
- Wootton, J & Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Computers chemistry.* 17:149-163.
- Ycas, M. (1974) *J. Theor. Biol.* 44: 145-160.
- Yusupov, MM; Yusupova, GZh; Baucom, A; Lieberman, K; Earnest, TN; Cate, JHD & Noller, HF. (2001). Crystal structure of the ribosome at 5.5 Å resolution. *Science.* 292: 883-292.

Páginas de Internet

- Servidor del KEGG <http://www.genome.ad.jp/kegg/kegg2>
- Mapa de Rutas Metabólicas en el KEGG <http://www.genome.ad.jp/kegg/pathway>
- PDB <http://www.rcsb.org/pdb/>
- Aplicaciones en el EMBOSS <http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/index.html>
- PHDsec <http://maple.bioc.columbia.edu/predictprotein/>
- Tcofee <http://igs-server.cnrs-mrs.fr/Tcofee/tcofee.cgi/index.cgi>
- Pfaat <http://pfaat.sourceforge.net/>

Los códigos de los programas del Apéndice 1 y las listas completas de los genes que presentaron cada tipo de mutación con su descripción metabólica para las tres enterobacterias pueden obtenerse en <http://bacteria.fciencias.unam.mx/Scripts/tesisD.htm>

APÉNDICE 1. TABLA DE PROGRAMAS MÁS IMPORTANTES ESCRITOS EN PERL

PROGRAMA	FUNCION
PD.pl	Realiza el método PD
Basuras.pl	Lee la salida del seg-1, analiza los datos y regresa una lista de los códigos de genes y sus respectivas secuencias simples.
Bella.pl	Ordena los resultados del Blastp de ortólogos.
B-num.pl	Da conteo de genes por vía metabólica.
Cai.pl	Calcula el CAI y decide quién es GTH usando Ds.
ConteoFaa.pl	Da porcentaje de SSR y no-SSR
Creador_uscod.pl	Genera una tabla de uso de codones a partir de una lista de genes
Cuenta.pl	Dice que lista de genes se encuentran en tal otra
Depura2.pl	Quita las parejas reiteradas
Estratagema_D.pl	Calcula el número de cambios necesarios para que una secuencia sobrepase las 4 Ds.
Estratagema_M.pl	Calcula el número de cambios necesarios para que una secuencia obtenga una mutación.
Genes_aledaños.pl	Identifica genes adyacentes.
Idefix.pl	Comprara las listas y da parejas bidireccionales (ortólogos).
Indice5.pl	Calcula el porcentaje de GC3 y decide quién es GTH usando Ds.
Kegg.pl	Hace amigable el formato del Kegg
Lista-lista.pl	Lee dos listas y de una quita la otra.
Localización2.pl	Especifica en dónde se encuentran las SSR dentro de los genes.
Mapa-par.pl	Grupos de parálogos que se mueven juntos
MarkovIN3D y 4D	Markov interactivo a diferentes Ds
Marsupilami.pl	Busca parálogos usando Blastp
Nodos.pl	Genes en nodos con mutaciones
Nombres.pl	Da los nombres de los genes y los ordena en bloques
Orto.pl	Hace los Blastp. $X_i \rightarrow Y_i$ & $Y_i \rightarrow X_i$.
Perdida.pl	Calcula pérdidas.
Periferie.pl	Da los cuatro genes aledaños por cada uno de la lista
Quito_parejas_redundates.pl	Elimina las parejas dobles
Seg.pl	Corre automáticamente el programa SEG dado un genoma con las diferentes salidas posibles.
Somospistolas.pl	Hace las tablas.
TODO2.pl	Los tres métodos de detección de GTH (CAI, GC3 y Markov). Obtención de 100 genes aleatorios de 30 genomas.
Uni-aa.pl	Cambia el formato de FASTA a sólo renglones.