UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

CENTRO DE CIENCIAS GENOMICAS

PROGRAMA DE GENOMICA COMPUTACIONAL

Uso de Codones, Traducibilidad, Niveles de Expresión y Transferencia Horizontal: ¿Hemos Sobreinterpretado Nuestros Organismos Modelo?

T E S I S

QUE PARA OBTENER EL GRADO DE
DOCTOR EN CIENCIAS BIOMEDICAS
P R E S E N T A:

Luis Arturo Medrano Soto

DIRECTOR DE TESIS:
Dr. Pedro Julio Collado Vides

Cuernavaca, Morelos **Junio de 2005**

Tutor Principal:

Dr. Pedro Julio Collado Vides
Centro de Ciencias Genómicas (CCG), UNAM.

Cotutor:

Dr. Gabriel Moreno Hagelsieb
Wilfrid Laurier University, Canada.

Cotutor:

Dr. Andrés Christen Gracia
Centro de Investigación en Matemáticas (CIMAT), Guanajuato.

Comité tutoral:

Dr. Pedro Julio Collado Vides
CCG–UNAM

Dr. Lorenzo Segovia Forcella
Instituto de Biotecnología (IBT), UNAM.

Dr. Jaime Mora Celis
CCG–UNAM

Miembros del Jurado:

Dr. Pedro Julio Collado Vides
CCG–UNAM

Dr. Enrique Merino Pérez
IBT–UNAM

Dr. Luis Eguiarte Fruns
Instituto de Ecología, UNAM.

Dr. José Andrés Christen Gracia
CIMAT

Dr. José Guillermo Dávila Ramos
CCG–UNAM

Dr. Alejandro Garciarrubio Granados
IBT–UNAM

Dr. Marco Antonio José Valenzuela
Instituto de Investigaciones Biomédicas, UNAM.

Reconocimientos

No hubiera sido posible realizar esta tesis doctoral sin el apoyo siempre generoso de mi tutor, el Dr. Julio Collado Vides, quien mediante su comprensión, motivación, orientación, paciencia y amistad siempre me instó a perseverar hasta alcanzar mis metas demostrándome que podía llegar más allá de mis propias expectativas.

Especialmente debo reconocer la contribución del Dr. Gabriel Moreno Hagelsieb, pues su cotutoría e infatigable disponibilidad a discutir el proyecto fue fundamental para su maduración y exitosa culminación. La asesoría del Dr. Andrés Christen Gracia permitió definir la metodología Bayesiana de modelación estadística en esta tesis y también realizar aportaciones en el área de minado de datos. Sin lugar a dudas, mi interacción con ambos doctores influyó mucho en mi formación. Agradezco la colaboración del Dr. Pablo Vinuesa, pues su apoyo permitió depurar las técnicas de análisis filogenético empleadas en la validación de los resultados.

Durante las etapas iniciales del doctorado fue esencial la orientación y retroalimentación que recibí por parte de mi comité tutoral, los doctores Julio Collado, Jaime Mora y Lorenzo Segovia. Sus sugerencias, siempre muy oportunas, canalizaron mi energía e interés sembrando la semilla que permitió la gestación de este proyecto.

Agradezco los valiosos comentarios a las distintas versiones del artículo, donde se reportan los resultados de este proyecto, por parte del Dr. Enrique Morett, Dr. Alejandro Garciarrubio, Dr. Enrique Merino, Dr. Warren F. Lamboy y el Dr. León P. Marínez–Castilla. Todos ellos enriquecieron la calidad científica de este trabajo. Doy un reconocimiento especial a todos mis compañeros de laboratorio, fue gracias a su apoyo, amistad, confianza, y carisma que el ambiente de trabajo siempre fue el ideal para desempeñar cualquier labor por ardua que fuera.

Mi gratitud es total para el Centro de Ciencias Genómicas, la UNAM y CONACYT, por abrirme sus puertas, dándome todo lo necesario para culminar mis estudios de la mejor manera.

Dedico esta tesis doctoral especialmente a ti madre, Luz María Soto Ceniceros, por tu amor incondicional, por todos los sacrificios que realizaste para educarme sin importarte las consecuencias, y por enseñarme, mediante el ejemplo, a no rendirme jamás. Es mucha la paciencia que has tenido para ver llegar este momento, pero finalmente aquí está, con todo mi corazón...

A mis hermanas, Alicia y Lucero, porque siempre están conmigo. A mis cuñados Ramón y Pepe por amar profundamente a mis hermanas. A todos mis sobrinos que quiero tanto: Hernán, Josué, Joel, Aleny y al recién llegado Aarón.

A ti Patty por todos los momentos que hemos vivido juntos...

A toda mi familia con sus dos reinas, mis abuelitas Pepa y Lola, por su confianza en mí y por todo el cariño que siempre me han prodigado.

Contenido

Resumen

Esta tesis presenta una evaluación al paradigma composicional para la predicción de genes transferidos horizontalmente, el cual postula que los genes foráneos muestran una composición atípica de codones en el genoma receptor al momento de ser adquiridos. Primero, se aplicaron varios criterios, incluyendo una validación filogenética, para identificar pares de genes exportado–importado donde aún se conserva la huella composicional del DNA donador. Posteriormente, se determinó cuantos de los genes detectados muestran un uso de codones (UC) pobre, típico o rico. En contraste con el paradigma composicional, los resultados muestran que la mayoría de los genes extranjeros, al ser adquiridos, exhiben predominantemente un UC típico en el genoma hospedero. Tal compatibilidad entre el UC de genes foráneos y el genoma receptor puede ser un prerrequisito para que la selección natural pueda evaluar las ventajas selectivas de funciones importadas por la célula. De ser así, un UC atípico actuaría como una barrera importante contra la adquisición y posterior utilización de genes extranjeros. Además, si el UC de genes foráneos es compatible con el nuevo contexto genómico del organismos que los recibe, entonces, se ha sobreenfatizado el papel del mejoramiento (*amelioration*) del UC en la transferencia horizontal, pues dicho fenómeno afectaría sólo a una minoría de los genes.

Otra implicación importante de los resultados es que el nivel típico de UC define una zona de seguridad o tolerancia, donde los genes extranjeros pueden ser traducidos eficientemente —presumiblemente debido a su compatibilidad con la maquinaria de traducción del genoma receptor. En apoyo a esta hipótesis, se muestra que en *Escherichia coli* K12 el UC promedio del genoma (representativo del nivel típico de UC) correlaciona significativamente con las concentraciones disponibles de tRNA. Además, los genes que utilizan preferentemente los codones más abundantes en la célula correlacionan mejor con las abundancias de tRNA que el UC de los genes altamente expresados, indicando así que los supuestos actuales para predecir niveles de expresión, partiendo únicamente del uso de codones, no están bien fundamentados. Si bien los genes altamente expresados tienen generalmente un UC por encima del promedio y correlacionan bien con la disponibilidad de tRNA, son tantos los factores que afectan el nivel de expresión que no es posible afirmar que todos los genes que muestran estos atributos son también altamente expresados.

Abstract

This thesis presents an evaluation to the compositional paradigm for horizontal gene transfer (HGT) prediction, which posits that foreign genes display atypical codon usage (CU) within the recipient genome immediately upon introgression. First, we applied several criteria, including phylogenetic validation, to identify pairs of imported–exported genes that still preserve the compositional footprint of the donor DNA. Then, we estimated the number of detected genes showing poor, typical or rich CU. Contrasting with the compositional paradigm, our results indicate that most alien genes exhibit predominantly typical CU in the host genome at the moment of acquisition, suggesting that such CU compatibility between foreign genes and acceptor genomes is a prerequisite to assess the selective advantage of imported functions. Thus, atypical (poor) CU may represent a strong barrier against successful integration and utilization of acquired genes. Furthermore, if the CU of alien genes is compatible with the genomic context of the new host, then, the role of amelioration in HGT has been overemphasized since it would happen only in a small fraction of genes.

Another important implication of the results is that typical CU defines a safety or tolerance zone, wherein foreign genes can be efficiently translated —presumably due to their compatibility with the translational machinery of the recipient organism. We found additional evidence supporting this hypothesis. In *Escherichia coli* K12, the genomic codon frequencies (representative of typical CU) correlate significantly with tRNA concentrations. Furthermore, genes preferentially using the most abundant codons in the genome correlate better with the tRNA pool than the CU of highly expressed genes. This result entails that the underlying assumptions of current methodologies to predict expression levels, based on CU alone, are not well founded. Even though highly expressed genes often show higher than average CU and correlate well with tRNA availability, there are so many other factors affecting gene expression that genes exhibiting these attributes are not necessarily highly expressed.

Presentación

El proyecto doctoral fue motivado por el gran caudal de información que se genera como producto de los diversos proyectos genómicos en el mundo. Ahora es factible mirar hacia atrás y verificar si, bajo la luz de muchos más datos, continúan siendo vigentes los supuestos fundamentales o hipótesis de trabajo que surgieron cuando todavía no se había completado la secuencia del primer genoma, es decir, durante la era pre–genómica. En particular, se examina la capacidad de predicción de los supuestos que relacionan al uso de codones (UC) con la transferencia horizontal de genes (THG) y los niveles de expresión.

El Capítulo I describe el análisis que se realizó para evaluar el "paradigma composicional", el cual postula que los genes foráneos muestran principalmente una composición atípica de codones en el momento de ser adquiridos por el genoma receptor. Los resultados aquí reportados no sustentan a este paradigma. Se observa que la gran mayoría de las THGs, donde todavía se conserva la huella composicional del DNA donador, involucran genes foráneos que al momento de ser importados despliegan directamente un UC típico —aún cuando ciertamente existe una elevada probabilidad de que los genes foráneos lleguen al genoma receptor exhibiendo un UC pobre (como consecuencia de la gran variabilidad del UC entre diferentes organismos). De este hallazgo se desprenden al menos cuatro conclusiones. Primero, aún antes de realizarse el intercambio horizontal ya existía una compatibilidad entre la composición de codones de los genes foráneos y del organismo aceptor. Segundo, el nivel típico de UC constituye una zona de tolerancia o seguridad donde los genes foráneos pueden ser expresados adecuadamente por ser compatibles con la maquinaria de traducción del genoma destinatario. Presumiblemente, tal compatibilidad es un prerrequisito para que la selección natural pueda evaluar la ventaja selectiva de funciones importadas por la célula. Tercero, si los genes transferidos exitosamente llegan directamente con un UC típico o rico, entonces no es necesario someter su secuencia a un proceso de "mejoramiento" (*amelioration*) para que refleje las tendencias en UC del genoma receptor. De ser así, tal proceso de "mejoramiento" del UC ha sido sobreenfatizado en la literatura, pues sólo sucedería en una minoría de los casos. Cuarto, un UC pobre representa una barrera considerable contra la adquisición y utilización de genes foráneos pues la célula no podría traducirlos adecuadamente. Debido al conflicto entre los resultados aquí reportados y los supuestos

esenciales de las metodologías de predicción de genes foráneos basadas en el paradigma composicional, resultó extremadamente difícil publicar los hallazgos. Sin embargo, después de una historia de 6 meses que involucró 4 rechazos, sin argumentación convincente por parte de 4 revistas internacionales, y de tres ciclos de revisión en la revista *Molecular Biology and Evolution* (que duraron otros 12 meses), finalmente el trabajo fue aceptado. El artículo publicado [1] se adjunta en el Anexo I al final de la tesis. En la discusión, al final del Capítulo I, se describe una serie de evidencias teóricas y experimentales, publicadas recientemente, que soportan fuertemente las conclusiones obtenidas.

En el capítulo I se propone que un nivel típico de uso de codones representa una zona de seguridad o tolerancia, donde genes foráneos pueden ser traducidos adecuadamente por el genoma receptor. La hipótesis subyacente es que dicha zona refleja la compatibilidad del UC de genes foráneos con la maquinaria de traducción del organismo hospedero. Esta hipótesis no es trivial y requiere de evidencias más sólidas que la sustenten. Por consiguiente, en el Capítulo II se explora la compatibilidad tRNA–UC, para determinar si genes con un UC típico muestran una correspondencia notable con la concentración de tRNA. Efectivamente, como se esperaba, el UC genómico (UC_G) correlaciona muy bien con la concentración de tRNA, apoyando así la noción de una zona de tolerancia. Sin embargo, la correlación UC_G vs tRNA resulto ser también más alta que la mostrada por el UC de las proteínas ribosomales (PRs) vs tRNA. Esto es inesperado, porque las PRs son el modelo estándar actual del tipo de genes cuyo UC correlaciona óptimamente con la disponibilidad de tRNA para maximizar la eficiencia de la traducción y los niveles de expresión. Por este motivo, se decidió estudiar si el UC_G constituye una mejor referencia para medir la compatibilidad de los genes con la maquinaria de traducción de la célula. Los genes que utilizan preferentemente los codones más abundantes en el genoma muestran una correlación más elevada con la concentración de tRNA que los genes conocidos o predichos como altamente expresados —aquellos que utilizan preferentemente los mismos codones que las PRs— sugiriendo así que el UC_G es mejor referencia para medir que tan eficientemente se puede traducir un gene (traducibilidad). Aquí hay un conflicto, las metodologías actuales de predicción de niveles de expresión parten del supuesto de que el UC en genes altamente expresados (e.g. las PRs) es óptimo para la traducción, implicando mayor correspondencia con la concentración de tRNA en comparación con genes de menor expresión, entonces ¿a qué se debe que existan genes no considerados como de alta expresión que muestran correlaciones más elevadas con la

abundancia de tRNA? El conjunto de análisis presentado en el Capítulo II representa una evaluación a los supuestos de trabajo actualmente empleados para predecir niveles de expresión. Lo resultados sugieren que tales supuestos no están bien fundamentados e involucran argumentos circulares. Se concluye que el UC está más relacionado con la eficiencia de la traducción que con el nivel de expresión. Por lo tanto, no es posible predecir confiablemente el nivel de expresión partiendo únicamente del UC. Los genes predichos como altamente expresados son sólo un subconjunto del total de genes traducibles eficientemente, pero no se puede afirmar que sean los más "óptimos" para la traducción. Se está trabajando en el manuscrito para publicar la contribución del Cápitulo II.

Durante el desarrollo del doctorado se trabajó de manera paralela en otro proyecto independiente al tema de tesis: el desarrollo de un método de clasificación Bayesiana (BClass por sus siglas en ingles *Bayesian Classifier*) que permite analizar datos biológicos de naturaleza heterogénea. Normalmente se utilizan métodos de agrupamiento (*clustering*) para realizar filogenias moleculares o estudiar patrones de expresión en microarreglos, porque los datos involucrados son matemáticamente homogéneos (tienen las mismas unidades) y el concepto de distancia entre los datos es fácilmente interpretable —las filogenias involucran distancias genéticas y los microarreglos diferencias en intensidades de expresión. Por otro lado, si se desea relacionar genes mediante un análisis que integre el nivel de expresión, vecindad en el cromosoma, la función molecular, el modo de regulación y la fuerza de los promotores, por citar un ejemplo, es común realizar varios análisis por separado, porque el concepto de distancia entre datos tan heterogéneos no tiene una interpretación útil. BClass permite realizar un análisis simultáneo de todas estas variables, mediante la transformación del conjunto de atributos biológicos heterogéneos en probabilidades de pertenencia a diferentes grupos. La transformación se logra al modelar cada variable biológica con una distribución estadística (i.e. Normal, Poisson, Multinomial, etc.) y después aplicar la teoría de modelos mezcla para calcular la probabilidad *a posteriori* de que cada entidad biológica (en este ejemplo genes) pertenezca a cada uno de los grupos en la mezcla. Este procedimiento elimina la necesidad indeseable de definir medidas de distancia o similitud para relacionar los genes. Al final, todos aquellos genes que muestren probabilidades similares de pertenencia a todos los grupos estarán relacionados. El artículo detallando esta metodología y su uso potencial [2] se encuentra adjunto en el Anexo II.

Capítulo I

Los genes importados exitosamente muestran un uso de codones *típico* en el genoma receptor al momento de ser adquiridos

Déjame decirte el secreto que me ha llevado a alcanzar mi meta. Mi fuerza reside exclusivamente en mi tenacidad.

LOUIS PASTEUR

1.1 Resumen del capítulo

El estudio de la *transferencia horizontal de genes*[1] (THG) ha despertado un gran interés por entender los mecanismos biológicos involucrados, sus implicaciones en la adaptación a un medio ambiente cambiante y su impacto en la evolución de las especies. Naturalmente, un problema esencial para alcanzar estas metas es la identificación confiable de genes que han participado en eventos de THG. Actualmente las metodologías teóricas para detectar genes que se han movido lateralmente pueden clasificarse en dos tipos: filogenéticas y composicionales. Los métodos filogenéticos, aunque no siempre es posible aplicarlos, cuentan con fundamentos más robustos y gozan de mayor aceptación. Sin embargo, si no se aplican con las debidas precauciones pueden arrojar resultados incorrectos; por ejemplo, al confundir genes *parálogos*[2] por *ortólogos*[3], o bien

[1] El intercambio de material genético (i.e. genes) entre especies diferentes.
[2] Genes que divergen después de un evento de duplicación genética dentro de un genoma. Tienden a adquirir nuevas funciones durante el curso de la evolución y suelen estar sujetos a diferentes presiones selectivas (ver Figura 1.1).

al analizar genes con tasas muy desiguales de mutación. Por otro lado, los métodos composicionales se pueden aplicar con mayor facilidad por no requerir la comparación de genes entre múltiples organismos, pero sus fundamentos teóricos e implicaciones respectivas son más debatibles. Los métodos composicionales consideran que los genes recientemente adquiridos por transferencia horizontal exhiben características atípicas en su secuencia de DNA, como el contenido de G+C, frecuencias de dinucleótios y uso de codones (UC); donde por atipicidad se quiere decir frecuencias significativamente diferentes al promedio genómico.

Con el propósito de evaluar los supuestos subyacentes y la capacidad de predicción de los métodos composicionales, este capítulo se concentra en determinar cuál es el nivel de UC (pobre, típico o rico) de los genes foráneos en el momento mismo de la transferencia. La teoría actual dicta que los genes importados exhiben predominantemente un UC "pobre", implicando que deben ser ineficientemente traducidos por la maquinaria del organismo receptor. La premisa fundamental en la estrategia para atacar esta incógnita plantea que en el instante en que se da el intercambio lateral, dos genes *xenólogos*[4] (ver Figura 1.1) son idénticos y por lo tanto guardan las mismas características composicionales, independientemente de si el UC es típico o atípico con respecto al genoma receptor. Como consecuencia, se asume que los genes extranjeros que aún conservan la huella composicional del DNA donador deben exhibir: (1) un UC muy similar; (2) aproximadamente la misma longitud; (3) la más alta similitud global a nivel de proteína, satisfaciendo por ende los criterios operativos actuales para reconocer ortología; y (4) su relación filogenética es irreconciliable con el árbol canónico de las especies. Los pares de genes que satisfacen las 4 condiciones son denominados Genes Candidatos a ser Xenólogos (GCXs).

Una vez identificados todos los pares de GCXs entre 103 genomas procariotes no redundantes, se comparó su nivel de UC con los niveles esperados por los métodos composicionales. Los resultados indican que la abrumadora mayoría de los GCXs despliegan un UC preferentemente típico en el genoma receptor al momento de la transferencia, derivándose así las siguientes conclusiones. Primero, un nivel típico de UC es un prerrequisito importante para que la selección natural pueda evaluar la ventaja selectiva de funciones importadas por la célula; segundo, el nivel típico de UC constituye una zona de seguridad o tolerancia donde los genes

[3] Genes en diferentes especies que evolucionaron del mismo gene ancestral a partir de un evento de especiación. Normalmente los genes ortólogos retienen la misma función en el curso de la evolución (Figura 1.1).
[4] Relación que surge cuando se intercambia material genético (e.g. genes) entre diferentes especies. El gene exportado (donado) y el gene importado (adquirido) están vinculados por una relación de xenología. (Figura 1.1)

extranjeros pueden ser expresados adecuadamente —presumiblemente debido a su compatibilidad con la maquinaria de traducción del genoma receptor; tercero, un UC pobre representa una barrera importante contra la adquisición y utilización de genes foráneos; cuarto, el papel del mejoramiento del UC, o "*amelioration*", en la transferencia horizontal ha sido sobreenfatizado, pues solo sucedería en una minoría de los genes. Aunque en aparente contradicción con los supuestos actuales, esta interpretación encuentra soporte en diversas evidencias teóricas y experimentales publicadas recientemente.

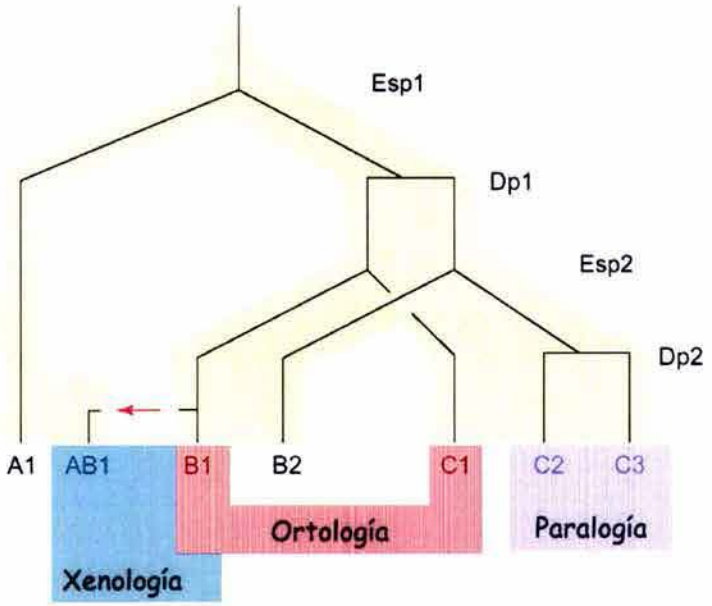

Figura 1.1. Tres tipos de Homología: Ortología, paralogía y xenología. Se muestra la evolución idealizada de un gene (líneas negras) a partir de un ancestro común, descendiendo hacia 3 poblaciones A, B y C (fondo amarillo claro). Hay dos eventos de especiación (Esp1 y Esp2) en los puntos donde se forman las "Y" invertidas. También hay dos eventos de duplicación genética (Dp1 y Dp2) ilustrados como líneas horizontales. Dos genes cuyo ancestro común reside en la unión de una "Y" invertida son ortólogos (e.g B1 y C1). Dos genes cuyo ancestro común reside en una línea horizontal son parálogos (e.g. C2 y C3). La flecha roja denota la transferencia del gene B1 de la especie B hacia la especie A. Aunque estrictamente hablando AB1 es xenólogo de los otros 6 genes, en este capítulo se relacionará con el terminó xenólogos, como definición de trabajo, al par de genes donado–adquirido (en este caso AB1 y B1). Los 7 genes son homólogos entre sí porque proceden de un mismo ancestro común en la raíz del árbol. Estas definiciones y el diagrama fueron tomadas del trabajo publicado por Walter M. Fitch [3].

1.2 Objetivo

Determinar cual es nivel de uso de codones (pobre, típico o rico) de los genes foráneos inmediatamente después de ser importados. Aclarar esta incógnita permitirá evaluar la generalidad del *paradigma composicional* para la detección de adquisiciones laterales recientes —genes transferidos horizontalmente muestran predominantemente una composición atípica de codones en el genoma receptor.

1.3 Antecedentes

La era de la secuenciación a gran escala y de los sistemas automatizados de anotación de genomas han generado bases de datos enormes a partir de las cuales se han realizado muchos descubrimientos. Análisis comparativos a nivel de DNA y de aminoácidos han revelado regiones aisladas o mosaicos de secuencia "atípica" altamente conservados, inspirando como resultado las preguntas de si estas secuencias fueron introducidas por transferencia horizontal o si son en realidad ocurrencias fortuitas que fueron exitosas y preservadas por selección natural.

El concepto de transferencia horizontal de genes (THG) involucrando orgánulos de eucariotes tiene una larga historia. A principios del siglo XX se propuso que los cloroplastos y las mitocondrias eran endosimbiontes bacterianos [4, 5]. Concepto que fue retomado y desarrollado cerca de 50 años más tarde [6]. Hoy en día ésta es una de las formas más aceptadas de movimiento horizontal a través de grandes barreras filogenéticas. El trabajo de Woese [7] demostrando que el rRNA mitocondrial y de cloroplastos está más relacionado con las bacterias que con eucariotes, ha representado la evidencia más convincente para la teoría de la endosimbiosis.

La era del DNA recombinante proporcionó información valiosa sobre el grado de conservación de los mecanismos genéticos y permitió demostrar experimentalmente que los genes pueden moverse a través de fronteras entre especies. Un muy buen ejemplo de THG que ocurre continuamente en la naturaleza es la transferencia natural de DNA plasmídico de la bacteria *Agrobacterium tumefaciens* a células de plantas, que resulta en la integración del DNA foráneo en el cromosoma de la planta, seguido por su expresión para generar cambios fenotípicos. Ciertamente, se sabía bien que los virus eran capaces de mediar la transferencia

horizontal mucho antes que el caso de *A. tumefaciens*. Aunque este fue un gran descubrimiento realizado mucho antes de los años 1960s (ver referencias en [8-10]), la transferencia horizontal entre microbios no tuvo el impacto que tuvo la transferencia entre microbios y eucariotes. La actual era genómica brinda oportunidades para explorar sistemas de THG que puedan existir entre diversos organismos.

El primer experimento que ilustró la habilidad del flujo de información genética entre especies pasó en gran medida desapercibido. En 1959 se descubrieron los plásmidos que transmiten resistencia a antibióticos, cuyo atributo era que contenían genes capaces de transmitir resistencia a múltiples antibióticos y que se transferían a través de diferentes especies bacterianas, demostrándose así que la información genética puede fluir de una especie a otra [11, 12]. Las implicaciones de este descubrimiento tuvieron un impacto profundo tanto en el campo de la ingeniería genética como en la teoría de evolución. Los primeros artículos que exploraron las implicaciones teóricas más profundas de la THG comenzaron a aparecer en los años 70s, aunque no fueron ampliamente reconocidos o aceptados. Por ejemplo, se observó que existen rasgos similares en plantas no relacionadas, pero que comparten el mismo ecosistema [13, 14], bajo este contexto se propuso que las plantas estaban intercambiando genes y se citó a la transferencia de genes plasmídicos como precedente de este tipo de eventos. También se planteó que la THG podría afectar la evolución en el reino animal [15, 16], e inclusive jugar un papel importante en la especiación [17].

Mientras tanto, los experimentos en ingeniería genética comenzaban a producir resultados sorprendentes. Por ejemplo, se introdujo un gene de levadura en una mutante de *Escherichia coli* deficiente en histidina, que resultó en el restablecimiento de la biosíntesis de histidina [18]. Lo que hoy en día es práctica rutinaria, era difícil de comprender a mediados de los 70s —genes de organismos eucarióticos artificialmente introducidos en bacterias podían en efecto funcionar. En 1980 se demostró que genes bacterianos podían expresarse exitosamente en levadura [19]. En 1983 se produjo el primer ratón transgénico que expresó un gene foráneo, el gene que codifica la hormona del crecimiento humano [20]. Diferentes experimentos demostraron, resultado tras resultado, que en el laboratorio se podían transferir genes entre especies y observar sus fenotipos. Las preguntas fundamentales que permanecieron fueron si estos eventos ocurrían efectivamente en la naturaleza y si sucedían en frecuencias suficientemente elevadas como para tener un impacto significativo en la evolución. En 1985 se propusieron dos explicaciones en apoyo a la

importancia de la THG [21]. Primero, si existían mecanismos tan potencialmente útiles de THGs a nivel molecular, la naturaleza debía encontrar una manera de utilizarlos. Segundo, una teoría evolutiva general que incorporara la idea del flujo de información genética a través de fronteras taxonómicas parecía proporcionar una respuesta simple y satisfactoria a la pregunta: ¿Por qué la biología molecular de todos los organismos vivos está tan unificada? Aun cuando los organismos pueden divergir independientemente después de la especiación, la biología ha retenido una unidad tan profunda que animales transgénicos pueden ser creados en el laboratorio.

Un factor adicional en favor de la relevancia de la THG surgió como producto del crecimiento de las bases de datos de ácidos nucleicos. A principios de los 80s ya se había acumulado para algunos *organismos modelo*[5] (i.e. *E. coli* y levadura) una muestra representativa de genes, lo cual permitió estudiar características composicionales del genoma y correlacionarlas con propiedades fisiológicas. Como resultado, se descubrió la relación que existe entre el uso de codones (UC), la concentración de tRNA y el nivel de expresión. En breve, las frecuencias de codones en un organismo no son azarosas, la mayoría de los genes siguen en mayor o menor grado las tendencias genómicas de UC [22]; además, los genes altamente expresados muestran un mayor *sesgo de UC*[6] que correlaciona significativamente con las especies de tRNA más abundantes [23, 24]. Estos hallazgos, junto con la demostración posterior de que un UC pobre puede afectar la eficiencia de la traducción [25-27], motivaron la proposición de dos ideas clave. Primero, genes con una composición atípica tanto de codones como de G+C podrían ser adquisiciones horizontales recientes [28]. Segundo, el nivel de expresión de genes heterólogos puede ser afectado por el grado de correspondencia entre el patrón de UC del gene introducido y el perfil preferido por el genoma receptor, por lo tanto se enfatizó la importancia biotecnológica de determinar un patrón de UC que promueva una expresión óptima [29]. En apoyo a estas ideas, se observó que genes de plásmidos y fagos no se apegan al UC genómico tan bien como genes cromosomales, llevando como consecuencia a la sugerencia de métodos generales de predicción de genes foráneos basados únicamente en la secuencia [30]. El razonamiento subyacente se basa en la hipótesis de que el UC refleja la adaptación de los genes nativos a la maquinaria de traducción de su genoma [23], y como los genes foráneos no han estado expuestos a las mismas

[5] Especies que son extensivamente estudiadas para comprender fenómenos biológicos particulares, esperando que los descubrimientos hechos en un organismo modelo podrán explicar como funcionan otros organismos. Esto funciona porque la evolución reutiliza principios biológicos fundamentales y conserva vías metabólicas, estrategias de regulación y mecanismos del desarrollo.

[6] Tendencia en los genes a usar un solo codón sinónimo por aminoácido.

presiones mutacionales y selectivas que los genes nativos, no es descabellado asumir que los genes foráneos deben exhibir una composición de codones pobremente adaptada al genoma receptor [30]. Este hecho señaló el nacimiento del *paradigma composicional* para la detección de THG cuando todavía faltaba casi una década para que se obtuviera la secuencia completa del primer genoma. Sin embargo, evaluaciones recientes de los métodos que se basan en este paradigma concluyen que son poco confiables si la composición atípica de secuencia se toma como única evidencia de la ocurrencia de THG [31-33].

En el terreno experimental, a mediados de los 80s ya se habían establecido varios mecanismos que mediaban el intercambio de genes, no sólo entre organismos unicelulares sino también entre metazoarios, promoviendo que muchos fenómenos biológicos difíciles de explicar se manejaran fácilmente haciendo alusión a la transferencia horizontal. Sin embargo, hubo una pausa en las observaciones que proporcionaban soporte directo a tales especulaciones. Con la secuenciación de genomas la situación ha cambiado. Actualmente, investigadores de áreas muy diversas están haciendo observaciones relacionadas con la THG. Como resultado, tal acumulación de evidencias hace factible buscar respuestas a preguntas como: (1) ¿Qué tan universales son los mecanismos de THG? y ¿Operan estos mecanismos en ambientes naturales? (2) ¿Cuál es la evidencia a favor de que la THG contribuye a los genotipos actuales de las especies? La evidencia principal a favor de que la THG es substancialmente común involucra un razonamiento filogenético. Sin embargo, hay dos problemas muy recurrentes en este tópico —determinar la topología real de un árbol de genes y la estimación de tiempos de divergencia. (3) Si los mecanismos existen y los eventos pueden documentarse, ¿juega la transferencia horizontal un papel significativo en la evolución? O bien, ¿Puede una teoría que incorpora DNA migratorio explicar fenómenos biológicos más generales?

A la fecha se han reportado numerosos casos de THG evidenciando que, en efecto, se trata de un fenómeno común [34-40]. Sin embargo, la propuesta de que la THG ha sido tan exhaustiva que elimina la posibilidad de describir la historia evolutiva de las especies mediante un árbol [41], ha sido impugnada de manera convincente por análisis colectivos de genes que soportan la existencia de tres dominios monofiléticos separados [39, 42, 43]. De hecho, se han acumulado evidencias sólidas indicando que el flujo horizontal de genes es mucho mayor al interior de linajes que entre linajes [39, 44-46]; por consiguiente, continua siendo razonable la idea de la existencia de una señal filogenética y de que un modelo jerárquico es adecuado para describir la

historia de las especies —pero es necesario recordar que la ausencia de filogenias discordantes no excluye la posibilidad de THG y que algunas especies pueden ser más susceptibles a la THG que otras [47]. Además, no todos los genes tienen la misma posibilidad de ser transferidos. La hipótesis de complejidad postula que es poco probable que los genes informacionales (aquellos involucrados en la transcripción, traducción y procesos relacionados) se transfieran en comparación a los genes operacionales (aquellos involucrados en el mantenimiento de la célula), debido a que típicamente requieren establecer más interacciones físicas con otros genes [48].

En resumen, se pueden distinguir dos tendencias en el estudio de la transferencia horizontal. Por un lado, se considera que la THG es un factor esencial en la evolución, capaz de dirigir la veloz adaptación a nuevos nichos y de inducir eventos de especiación [34, 41, 49-51]. Esto se debe a que en principio es mucho más rápido importar genes ya fabricados y listos para responder a retos ambientales que experimentar con secuencias nativas. Por otro lado, se argumenta que tal interpretación es una exageración propiciada, en parte, por confiar en métodos inadecuados para la identificación de eventos de THG. Aunque la THG puede ser frecuente, la fijación de secuencias foráneas en poblaciones es poco probable, porque la gran mayoría de las secuencias adquiridas lateralmente no le otorgan al genoma receptor una ventaja selectiva. Por lo tanto, el impacto de la THG en la evolución de los genomas bien puede ser marginal [52-54]. Hay una gran cantidad de ejemplos donde mutaciones simples afectan la traducción y disminuyen la velocidad de crecimiento en ausencia de una condición de selección que las compense [55]; es natural esperar que algo similar suceda con secuencias foráneas. En una situación estable, es muy probable que las mutaciones sean destructivas o neutrales y la probabilidad a priori de fijación de una secuencia neutral es inversamente proporcional al tamaño de la población [56]. Es decir, una vez que un linaje celular ha evolucionado componentes bien integrados, es muy poco probable que componentes mutantes o foráneos incrementen la viabilidad del linaje. Las adquisiciones neutrales se difundirán en la población, pero también serán blanco de mutaciones aleatorias y erradicadas por deriva genética.

A continuación se presenta una discusión sobre las cualidades, defectos y sesgos de los métodos actuales para identificar transferencias horizontales. Esto es fundamental porque dilucidar si el nivel de UC de los genes foráneos es típico o atípico con respecto el genoma receptor, en el momento de la transferencia, requiere de la detección confiable de pares de genes donador/receptor involucrados en eventos de THG.

1.3.1 Métodos de detección de transferencia horizontal.

Determinar si la THG es o no un fenómeno frecuente en la naturaleza, plantea el problema teórico de identificar cuando un gene o región de DNA se ha originado a partir de un movimiento horizontal. Los métodos que han surgido pueden clasificarse en dos grandes categorías: aquellos basados en criterios filogenéticos y aquellos basados en propiedades composicionales de la secuencia.

1.3.1.1 El método de incongruencia filogenética

Este método es el más confiable para detectar la ocurrencia de transferencias horizontales de genes. Consiste en tomar un grupo de genes ortólogos (ver Figura 1.1) pertenecientes a un conjunto de especies razonablemente lejanas, para luego construir un árbol filogenético y compararlo con la filogenia conocida de esas especies. Si se observa una incongruencia entre el "árbol de genes" y el "árbol de las especies", entonces se puede plantear un posible caso de transferencia horizontal. Esta prueba se ha aplicado con varios grados de rigor desde las primeras afirmaciones de transferencia horizontal y ha sido descrita con mucho detalle en la literatura [57]. Entre las primeras aplicaciones de este criterio se encuentra la presentada por Woese y Fox en 1977 [58].

Idealmente la aplicación del método de incongruencia filogenética requiere que se satisfagan varias condiciones. Primero, los genes bajo análisis deben contener información filogenética. Segundo, los genes comparados deben ser ortólogos y no parálogos (ver Figura 1.1). Finalmente, el ejemplo de incongruencia debe involucrar un gene cuya tasa de sustitución no sea radicalmente diferente a la de los otros genes que se están comparando. Para que estas 3 condiciones puedan garantizarse, el número de genes a analizar debe ser razonablemente grande, i.e. más de 5 y posiblemente más de 10 [59], minimizando así errores debido al muestreo. Es difícil dilucidar la dirección de las transferencias a partir de incongruencias filogenéticas, especialmente para transferencias ancestrales que involucran linajes que dieron origen a muchas especies actuales. Por ejemplo, Doolittle y colegas [60] concluyeron que la enzima gliceraldeido 3-fosfato deshidrogenasa (*gapdhA*) en *E. coli* fue adquirida horizontalmente desde un eucariote, dado que era el único procariote presente en un clado de eucariotes. Sin embargo, con el hallazgo posterior de un ortólogo de *gapdhA* en *Anabaena* parece ahora más probable que una bacteria ancestral donó este gene a los eucariotes [61].

1.3.1.1.1 Problemas debido a paralogía

Muchos de los reportes prematuros de posibles THG fueron producto de la comparación de genes parálogos al ser tratados como genes ortólogos. Los árboles construidos a partir de genes parálogos pueden ser incongruentes como consecuencia de divergencia funcional, involucrando así diferentes presiones selectivas y por lo tanto distintas tasas de sustitución. Este escenario puede darse también como producto de un simple error de muestreo; cuando se analizan muy pocos genes y/o sus productos no han sido totalmente caracterizados. El problema de paralogía fue responsable de que se infiriera que la enzima Cu-Zn superóxido dismutasa de la bacteria *Photobacterium leiognathi* proviniera de una fuente eucariótica y de que la leghemoglobina de plantas viniera de vertebrados. Conforme se analizaron más secuencias y se identificaron correctamente los genes ortólogos, se encontró que los árboles de los genes respectivos son en realidad razonablemente congruentes con el árbol de las especies [62, 63].

1.3.1.1.2 Problemas generados por tasas desiguales de mutación

Diferencias en tasas de substitución pueden no ser evidentes en conjuntos con pocos datos. Como lo notó Felsenstein [64], la comparación de genes que están sometidos a tasas muy desiguales de sustitución puede resultar en "afinidades" aberrantes durante la reconstrucción filogenética. Este problema se encontró en la calmodulina de músculo estriado de pollo. Gruskin y colaboradores [65] mostraron que el gene designado como tipo calmodulina (*cl*), era muy divergente del otro gene de calmodulina en el pollo (*cam*), así como de cualquier otro gene en vertebrados. Partiendo de este hecho, se sugirió que *cl* entró en el pollo por transferencia horizontal, posiblemente a partir de una retrotransposición mediada por virus (porque el gene no tiene intrones). Sin embargo, análisis posteriores [59] pusieron de manifiesto que el gene *cl* del pollo varía mucho más rápido que sus contrapartes en vertebrados, y además puede no ser ortólogo de los genes con los que originalmente se comparó. Por lo tanto, el gene no puede ser considerado como adquirido horizontalmente.

1.3.1.1.3 Problemas debido a convergencia evolutiva

Algunos científicos no aceptan la THG como única explicación para una incongruencia filogenética; en su lugar se propone la posibilidad convergencia evolutiva. Por ejemplo, Kemmerer y colaboradores [66] mostraron que el citocromo c en Arabidopsis es similar al citocromo de hongos, pero no ofrecieron una explicación mecanística. Posteriormente, en 1994,

Doolittle al hacer una revisión de este problema concluyó que, aunque la convergencia mecanística–funcional es común y la convergencia estructural enzimática probablemente ha ocurrido, no se había establecido a la fecha un caso genuino suficientemente convincente de convergencia de secuencia [67]. En un caso ampliamente citado de convergencia evolutiva, el de la lisozima de langur convergiendo hacia la de rumiantes [68], Doolittle mostró que el árbol de la lisozima es congruente con el árbol de las especies. Esto es, las substituciones convergentes de aminoácidos que pudieron ocurrir en el linaje que lleva a los rumiantes y al langur fueron pocas en el trasfondo de cambios neutrales como para ocultar la afinidad de la lisozima del langur con la de los primates.

1.3.1.1.4 Pruebas de significación estadística

Toda conclusión que involucre el hallazgo de una relación "inesperada" a partir del método de incongruencia filogenética, requiere que se estime la confianza estadística del resultado, permitiendo así evaluar si la observación "inesperada" es significativa. Desafortunadamente, en general no hay pruebas estadísticas suficientemente rigurosas para determinar la confiabilidad de árboles filogenéticos. Los problemas computacionales son inmensos [69]. Por ejemplo, para conjuntos de datos que involucren una gran cantidad de especies, puede ser extremadamente difícil encontrar inclusive el árbol más corto, sin mencionar la prueba de confianza de ese árbol contra algún otro. Este problema ha atraído mucha atención y se han propuesto métodos para calcular la confiabilidad de árboles para varias especies. Entre ellos están el método de máxima verosimilitud [70, 71] y el de máxima parsimonia [72]. El problema con el procedimiento de máxima verosimilitud es que antes de calcular la confiabilidad, debe asumirse un modelo evolutivo. Este modelo usualmente asume que los reemplazos a través de linajes y los eventos de ramificación siguen un proceso markoviano. Empero, una cosa es preguntar si un árbol particular es consistente con un modelo específico, y otra cosa muy diferente es preguntar si puede discriminar entre dos posibles modelos.

Construir un árbol filogenético a partir de un conjunto de datos que contiene "homoplasia" puede hacer la prueba de incongruencia filogenética aún más difícil. La homoplasia surge cuando especies evolutivamente lejanas comparten rasgos únicos. La dificultad yace en distinguir si los rasgos compartidos reflejan la herencia a partir de un ancestro en común, o si surgieron independientemente. Tradicionalmente se considera que la homoplasia es el resultado de

procesos tales como la convergencia y reversión a estados ancestrales; obviamente la THG también contribuiría a la homoplasia.

1.3.1.1.5 Elementos móviles

Si el método de incongruencia filogenética es aplicado en la ausencia de otra evidencia puede ser demasiado restrictivo —impediría que se consideren genes que están frecuentemente involucrados en THG. Este es probablemente el caso de muchos elementos transponibles porque su transferencia es tan frecuente que la filogenia de las especies que alojan estos elementos se pierde totalmente. Este problema ya se ha revisado con detalle en la literatura. Un ejemplo famoso de transferencias de genes eucarióticos, y a su vez uno de los casos mas convincentes, es el factor P de *Drosophila melanogaster* [73]. Este caso es persuasivo porque la transferencia ocurrió en años recientes y por lo tanto se observó cuando sucedió. Monitorear el evento en tiempo real en poblaciones naturales es muy convincente, tanto como observar la diseminación de genes resistentes a antibióticos por medio de plásmidos entre bacterias patogénicas. Además, en este caso, el método de congruencia filogenética apoya fuertemente la transferencia horizontal [74]. La historia filogenética de muchos elementos móviles se parece mucho a la filogenia de virus (puesto que algunos están mezclados con virus) en que su historia es relativamente independiente de la filogenia de sus respectivos hospederos [75].

1.3.1.1.6 La estructura mosaico en cromosomas

Hasta hace relativamente poco todavía se cuestionaba si la información genética podía fluir entre diferentes cepas de *E. coli*. En un inicio se argumentaba que el *flujo génico*[7] no podía ser significativo en *E. coli*, porque de ser así las diferencias entre cepas hubieran desaparecido. Sin embargo, el proceso que se siguió para contestar a esta pregunta ha conducido al desarrollo de nuevos criterios para detectar tanto flujo génico como transferencias horizontales entre especies filogenéticamente cercanas.

El grupo de Selander se enfocó en el análisis de poblaciones naturales de *E. coli* y concluyeron que la estructura de su población era "clonal", proponiendo por lo tanto que el flujo de genes y recombinación entre cepas naturales de *E. coli* no debe ser importante [76, 77]. Esta conclusión estaba basada en el hallazgo de que las poblaciones naturales de *E. coli* podían ser divididas, usando una distancia genética derivada de polimorfismos de enzimas, en al menos tres

[7] Transferencia de genes al interior de una especie.

grupos donde los miembros de un grupo estaban más estrechamente relacionados entre sí que con los miembros de otros grupos.

La noción de que no hay flujo génico entre cepas de *E. coli*, debido a la estructura clonal de su población, fue descartada después de que se realizaron comparaciones entre secuencias de mayor tamaño obtenidas de diversas cepas de *E. coli* [78-80]. Para hacer las comparaciones, se secuenció una región de 4400 pb del operon *trp* de 36 cepas de *E. coli*, seleccionadas del mismo conjunto de cepas que utilizó el grupo de Selander para determinar la estructura clonal de la población. Estos análisis confirmaron el hallazgo de que las 36 cepas podían ser divididas esencialmente en los mismos grupos obtenidos por los polimorfismos de enzimas. Sin embargo, también se encontró que cuando se comparaban cepas dentro de un grupo, uno de los miembros puede tener una sección corta que difiere de los otros miembros. Esto es, dentro de las regiones individuales de similitud se encuentran esparcidas regiones de disimilitud. Además, la región de disimilitud podía encontrarse a menudo en alguno de los otros grupos, como si esta región hubiera sido transferida de un grupo a otro. A partir del tamaño promedio de las regiones de disimilitud, se ha estimado que un evento promedio de recombinación resulta en la transferencia de algunos cientos o hasta miles de pares de bases. Se dice que los pares de cromosomas que siguen este patrón tienen una estructura mosaico. Este análisis demostró la presencia de subpoblaciones en *E. coli* que son genéticamente distintas, pero que ocasionalmente intercambian material genético sin destruir su identidad. En general, dos secuencias homólogas (ver Figura 1.1) de DNA que muestren un cambio abrupto de similitud, en una región bien delimitada, presentan la posibilidad de una estructura mosaico.

1.3.1.2 Métodos Composicionales

El progreso en la caracterización de diferentes cepas patogénicas de *Salmonella* ha llevado a numerosas propuestas de flujo génico. Hay muy buenos ejemplos involucrando factores virales. Por ejemplo, se ha mostrado que los antígenos de superficie utilizados para clasificar serotipos de *Salmonella* están distribuidos de forma discontinua a través de cepas lejanamente relacionadas [81], lo cual sugiere que los genes de estos serotipos se han movido dentro de esta especie. Groisman y colegas [82] han visto al cromosoma de *S. typhimurium* como mosaicos de partes distantemente relacionadas. Esta conclusión se basa, en parte, en la comparación de genes entre enterobacterias. Aún cuando *E. coli* y *S. typhimurium* comparten genomas de tamaño similar,

con el 90% de sus genes mostrando altos niveles de sintenia e identidad, cerca del 10% de los genes en *S. typhimurium* codifica funciones totalmente ausentes en *E. coli*. Además, el contenido de G+C en estos genes únicos con frecuencia es significativamente menor al promedio de todo el genoma; un hallazgo que aparentemente apoya la idea de un origen remoto de estos genes, aunque también se han propuesto hipótesis que rechazan tal posibilidad, argumentando que estas secuencias pueden ser nativas y estar sujetas a diversas presiones selectivas producto de su participación, directa o indirecta, en distintos procesos biológicos [59, 83]. De hecho, el planteamiento de posibles donadores remotos para algunos de estos genes —como *phoN* [84] y un regulador transcripcional [85]— es problemático porque la única evidencia del origen remoto es la desviación del contenido de G+C.

Debido a que diferentes factores pueden influir en el contenido de G+C, Syvanen examinó en 1994 la hipótesis de origen remoto con mayor detalle [59]. Al estimar la distribución del contenido de G+C a partir de 757 fragmentos de DNA de *E. coli* y 131 de *S. typhimurium*, Syvanen observó que la distribución puede dividirse en dos grupos. La mayoría de los fragmentos se agruparon alrededor de 0.509 de G+C (cerca del promedio genómico) con una distribución aproximadamente normal. La segunda clase mostró una desviación significativa hacia bajo G+C. Por el criterio de contenido de G+C, estos fragmentos son candidatos a transferencias horizontales. Sin embargo, hay un problema con este argumento, la desviación es únicamente hacia bajo contenido de G+C. La variación hacia alto contenido de G+C es consistente con la varianza predicha para fluctuaciones aleatorias de G+C dada una media de 0.509. ¿Por qué no se ven genes con alto contenido de G+C que vengan de fuentes remotas?

El caso de *S. typhimurium* no fue muy diferente. El principal agrupamiento cerca a la mediana de 0.516 es sólo aproximadamente normal (posiblemente debido al tamaño más pequeño de la muestra), pero la mayoría de la desviación es, como en el caso de *E. coli*, hacia un contenido de G+C bajo. Syvanen concluyó que debido a que la desviación se da principalmente hacia bajo G+C, es poco probable que esta sea una evidencia de origen remoto, argumentando que la selección funcional por bajo G+C es más viable. Obviamente estas regiones tendrían una temperatura de desnaturalización baja, y sería fácil imaginar escenarios donde mecanismos de replicación o recombinación permitieran seleccionar estas regiones. Por ejemplo, Syvanen propone que una explicación más simple para el bajo nivel de G+C es que se trata de DNA que

participa frecuentemente en rearreglos genómicos. El nivel más bajo de G+C pudo ser entonces seleccionado en el paso de recombinación al facilitar la desnaturalización del DNA.

1.3.1.2.1 El criterio de uso de codones

A mediados de los 80s empezaron a surgir métodos formales, no basados en análisis filogenéticos, con la intención de detectar genes de origen foráneo. Tales métodos se sustentan en la observación de que al interior de un organismo los genes tienden a seguir el patrón de UC del genoma [22], y por lo tanto aquellos genes que claramente se salían de este patrón fueron interpretados como adquisiciones horizontales recientes de origen remoto –el paradigma composicional. Este es un criterio difícil de aplicar porque el UC de proteínas pequeñas o poco abundantes se desvía del sesgo genómico. Médigue y colaboradores [86] examinaron el UC de 740 genes de *E. coli*, y encontraron tres clases de genes: (1) las proteínas altamente expresadas que definen el sesgo genómico; (2) las proteínas de expresión moderada que utilizan algunos codones raros; y (3) un grupo residual que muestra una marcada preferencia por codones raros. Este tercer grupo contiene la mayoría de los genes que serían predichos como nómadas, tales como secuencias de inserción y otros elementos móviles.

En organismos modelo los genes altamente expresados (e.g. proteínas ribosomales) muestran una composición de codones bien adaptada al genoma, y sus preferencias de codones sinónimos son consideradas como óptimas para maximizar la eficiencia de la traducción pues correlacionan bien con la concentración de tRNA [23, 24]. Dado que el sesgo en UC en genes de alta expresión es mayor al sesgo promedio del genoma, las metodologías que se desarrollaron a partir de entonces asumieron que todo gene con UC atípico (diferente tanto al promedio genómico como al UC de las proteínas ribosomales) fuera predicho como foráneo [30, 34, 36, 86-92]. Es importante mencionar, sin embargo, que hace una década ya se había recomendado precaución en el uso de estos criterios; si bien la exploración del UC es un ejercicio interesante en el análisis de secuencias, las explicaciones alternativas para cualquier desviación de la tendencia promedio de un organismo son suficientemente numerosas (e.g. rearreglos genómicos, mantenimiento de estructura secundaria, estabilidad, propiedades del DNA reflejadas en el mRNA como la susceptibilidad al daño mutagénico, señales relacionadas con la replicación, etc.) para impedir su uso como criterio único en la predicción de THG [59]. Además, evaluaciones

más recientes a estás metodologías indican que son poco confiables [31, 32]. En el mejor de los casos, el UC puede ser usado como apoyo a otras evidencias más sólidas.

1.4 Estrategia para determinar el nivel de UC de los genes foráneos en el momento de la transferencia

Dilucidar si la atipicidad del UC puede seguir siendo considerada como un detector confiable de THG, requiere del diseño de una metodología que en principio no favorezca genes con tendencias composicionales particulares. La premisa fundamental de trabajo en este capítulo plantea que inmediatamente después de una transferencia horizontal, el DNA donador es idéntico en secuencia y tamaño al DNA aceptor, independientemente de si los genes transferidos muestran un UC típico o atípico en el genoma receptor. El empleo del UC como parámetro en el estudio de la THG es relevante porque además de transmitir información valiosa sobre la composición de nucleótidos, también es un indicador del grado de compatibilidad entre genes individuales y la maquinaria de traducción de la célula. Entonces, es necesario identificar, primero, pares de genes xenólogos donde la huella composicional del DNA donador esté bien conservada y posteriormente preguntar si el UC es típico o atípico. Siguiendo este principio, se proponen cuatro condiciones básicas para identificar pares de genes candidatos a ser xenólogos (GCXs). Todo par de GCXs debe: (1) tener un UC muy similar; (2) mostrar aproximadamente el mismo tamaño; (3) sus secuencias de aminoácidos deben exhibir los niveles más altos de identidad global, cuando se comparan con las secuencias de sus probables ortólogos (POs) en otros organismos; y (4) la relación filogenética entre ellos debe ser irreconciliable con el árbol canónico de las especies. Sólo hasta después de haber obtenido un conjunto de GCXs que satisfagan las cuatro condiciones, se debe preguntar cuáles son las tendencias que exhiben en su nivel de UC y contrastarlas con los niveles que se esperarían tanto por los criterios de otras metodologías como por azar. La Figura 1.2 muestra la estrategia general que se siguió para predecir pares de genes xenólogos que tuvieran una composición similar de codones y posteriormente comparar su distribución de UC con el potencial de THG, es decir con los niveles de UC que se esperarían al azar dados los genomas analizados.

Una descripción científica detallada del trabajo se publicó recientemente [1] y se encuentra incluida en el Anexo I al final de la tesis. Para minimizar redundancias, aquí sólo se mencionarán las partes relevantes pero se profundizará en aquellos detalles que no se trataron en el artículo.

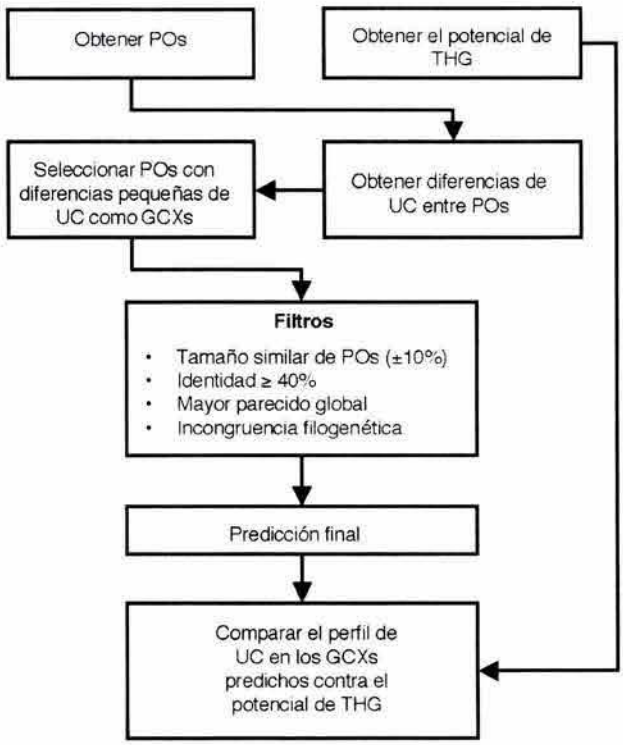

Figura 1.2. Estrategia para determinar el nivel de uso de codones (UC) de los genes foráneos en el genoma receptor en el momento de la transferencia. Primero, se estima el potencial de transferencia horizontal de genes (THG) entre todos los genomas (Sección 1.6), es decir, el nivel de UC que un gene cualquiera mostraría en otro genoma de ser transferido en este instante. Segundo, se obtienen todos los pares de posibles ortólogos (PO) entre los genomas analizados (ver Sección 1.7). Tercero, tomando un gene como referencia se calcula su diferencia de UC con todos los POs respectivos (Sección 1.8). Cuarto, se seleccionan aquellos casos donde el gene referencia muestra una diferencia muy pequeña de UC con algún PO y posteriormente son sometidos a varios filtros para predecir genes candidatos a ser xenólogos (GCXs); ver Sección 1.9. Quinto, finalmente se compara el nivel de UC de genes xenólogos con el potencial de THG (Sección 1.10).

Tomando en cuenta el cuerpo de evidencias que relaciona al UC con la eficiencia de la traducción, se diseño el índice de riqueza de codones (CRI por sus siglas en ingles *Codon Richness Index*), que cuantifica el grado en que los genes utilizan los codones más abundantes de un genoma referencia (ver Sección 1.5). El potencial de transferencia horizontal de genes (THG), o probabilidad al azar de que un gene foráneo despliegue un UC pobre, típico o rico en el genoma receptor si ocurriera una transferencia en este instante, se calcula como se describe en la Sección 1.6. Inmediatamente después de una THG los genes intercambiados son idénticos y cumplen, por

lo tanto, todos los criterios impuestos por los métodos actuales para detectar ortología con base en la secuencia de aminoácidos. Por este motivo, los GCXs se buscaron entre el conjunto de probables ortólogos (POs) que fueron identificados como se describe en la Sección 1.7. Otro punto fundamental es el método a utilizar para medir y comparar el UC. Se utilizó un enfoque Bayesiano para discriminar pares de GCXs que tienen un UC significativamente más similar entre sí que con otros POs relacionados, el método se detalla en la Secciones 1.8 y 1.11. Como un UC similar entre GCXs no es evidencia contundente de THG, se aplicaron otros criterios que incrementan sustancialmente la confianza en las predicciones, esto es, el tamaño similar de POs, máximo parecido global a nivel de aminoácidos y la incongruencia filogenética con el árbol universal de las especies (ver Sección 1.9). La comparación entre el potencial de THG más la discusión de las implicaciones biológicas y evolutivas de los resultados se presentan en la Sección 1.10 y en el artículo incluido al final de la tesis (Anexo I).

1.5 Una medida adecuada del uso de codones que refleje traducibilidad

Evaluar el nivel de UC no es trivial porque existe más de una alternativa para hacerlo, en todos los casos se requiere de un modelo para cuantificar, comparar e interpretar preferencias de codones sinónimos. Los supuestos que conforman la columna vertebral de las metodologías actuales surgieron durante la era pre–genómica y, aunque evidentemente razonables en su momento, datos recientes sugieren que fueron producto tanto de la escasez de datos como de una sobreinterpretación del conjunto limitado de evidencias experimentales disponibles en ese tiempo; el Capítulo II hace una revisión detallada de este problema, justificando la necesidad de desarrollar una nueva medida de UC que refleje de manera más adecuada la relación entre la composición de codones y la eficiencia de la traducción.

Los métodos más populares para estudiar el UC, no toman en cuenta la composición de aminoácidos, ni los codones que codifican para señales de paro, metionina y triptófano, debido a que se asume implícitamente, primero, que la composición de aminoácidos está más comprometida con la función que con el proceso traducción y por lo tanto su contribución es poca o nula a la *eficiencia de la traducción*[8] (traducibilidad) en comparación al UC. Segundo, que estos codones no son informativos en cuanto al proceso de traducción —o bien no codifican para

[8] La rapidez con la que el ribosoma traduce un RNA mensajero una vez que se ha tomado en cuenta la estabilidad del mRNA. Normalmente se mide como el número de proteínas traducidas a partir de un mRNA.

aminoácidos o no tienen codones sinónimos que compitan por una especie particular de tRNA. Sin embargo, se han reportado evidencias apoyando la noción de que el uso de aminoácidos sí tiene un impacto considerable en la traducibilidad de los genes. Por ejemplo, se ha observado que la composición de aminoácidos esta relacionada significativamente con el nivel de expresión [93], y que existe una tendencia a utilizar aminoácidos cuya biosíntesis es menos costosa, en términos energéticos, en genes que exhiben correlaciones UC–tRNA elevadas [94]. Es necesario aclarar que los codones para metionina y triptófano son traducidos por tRNAs que pueden estar en concentraciones mayores, o menores, que otros tRNAs traduciendo aminoácidos con varios codones sinónimos, y por lo tanto su contribución a la traducibilidad no es despreciable (ver Capítulo II, Secciones 2.5, 2.6 y 2.7). Además, todos los genomas muestran una marcada preferencia por alguno de los codones de paro; es posible que se deba a una estrategia para minimizar los costos de errores en la terminación de la traducción [95], incrementando como consecuencia las tasas de traducción porque la tasa de producción de proteína, a partir de mRNAs de un cierto tipo, es igual a la tasa de terminación de la traducción de esos mensajeros [96]. El sesgo en uso de codones de paro correlaciona bien con el hecho de que, en procariotes, los factores de terminación de la traducción reconocen estos codones con distintas afinidades [97]. También se ha propuesto que, tanto en procariotes como eucariotes, hay señales conservadas al final de los genes que pueden promover una terminación eficiente de la traducción, ya sea en la forma de tetra–núcleotidos [98, 99] o bien como interacciones directas entre el factor de terminación y el último peptidil-tRNA$^{Ser/Phe}$ [100]. Inclusive, en genes contiguos el sesgo en codones de paro también podría estar relacionado con presiones selectivas por evitar estructuras secundarias en los puntos donde termina un gene y empieza el otro [101]. Se deja para el Capítulo II la comparación y discusión completa de los problemas asociados a las metodologías actuales que evalúan el UC. Por el momento es suficiente decir que una medida más adecuada de traducibilidad debe tomar en cuenta las abundancias de todos los codones, sean degenerados o no, y la contribución de la composición de aminoácidos; entre más un gene utilice los codones más abundantes, mayor será su correlación con la disponibilidad de tRNAs, indicando que será traducido con mayor eficiencia.

Debido al conjunto de evidencias, expuestas en el párrafo anterior, indicando que el uso de aminoácidos más los codones que codifican para señales de paro, metionina y triptofano están relacionados con la traducibilidad de los genes, se diseño el Índice de Riqueza de Codones (CRI

por sus siglas en inglés *Codon Richness Index*) donde se toman en cuenta las frecuencias de los 64 codones para cuantificar el grado en que genes individuales utilizan los codones más abundantes en un genoma referencia (ver Sección 1.5.1). Si en general el UC es homogéneo como lo dicta la hipótesis del genoma [22], el sesgo de UC correlaciona con las abundancias de tRNA [23], las abundancias de aminoácidos influyen en la eficiencia de la traducción [93, 94] y la célula debe expresar todos sus genes a niveles adecuados para sobrevivir, entonces es posible interpretar al CRI como una medida de traducibilidad dado que toma en cuenta todos estos factores. El Capítulo II, Sección 2.6, presenta la comparación del CRI con otros índices y muestra como genes con alto CRI correlacionan mejor con las concentraciones de tRNA que los genes conocidos o predichos (por otros índices de UC) como altamente expresados; también se analizan otras evidencias que respaldan esta interpretación.

1.5.1 El índice de riqueza de codones (CRI)

En la sección 1.5 se dieron argumentos biológicos para exponer algunos problemas con las metodologías actuales que evalúan el uso de codones, y en el Capítulo II, Sección 2.6.2, se exponen las razones técnicas para no utilizar el índice estándar de uso de codones CAI (por sus siglas en ingles *Codon Adaptation Index* [30]).

Sea $G_{a,i}$ el gene i en el genoma a, $n_{a,i}(c)$ el número de veces que el codon c aparece en el gene $G_{a,i}$, y $L_{a,i}$ la longitud en codones (incluyendo el codón de término) del gene $G_{a,i}$, esto es $L_{a,i} = \sum_{c=1}^{64} n_{a,i}(c)$. La frecuencia relativa, $q_{a,i}(c)$, del codón c en el gene $G_{a,i}$ normalizada por $L_{a,i}$ es entonces definida como $q_{a,i}(c) = \frac{n_{a,i}(c)}{L_{a,i}}$ (se tiene que $\sum_{c=1}^{64} q_{a,i}(c) = 1$).

Ante el evento potencial de una transferencia horizontal del gene $G_{a,i}$ al genoma b, es posible estimar si el UC de $G_{a,i}$ es compatible a priori con el UC (o nuevo contexto genómico) del genoma receptor b. En términos más precisos, se trata de cuantificar el grado en que el gene $G_{a,i}$ usa los codones más abundantes de b. La frecuencia o abundancia genómica de cada codón c en el genoma b se puede interpretar como la probabilidad de encontrar ese codón en el conjunto total

de genes del genoma *b*, $p_b(c)$, y se calcula como $p_b(c) = \frac{N_b(c)}{\sum_{j=1}^{64} N_b(j)}$, donde $N_b(c)$ es el número total de veces que el codón *c* aparece en el genoma *b* (se cumple la condición $\sum_{c=1}^{64} p_b(c) = 1$).

Considerando a $p_b(c)$ como el peso o contribución del codon *c* a la distribución genómica de *b*, se puede ponderar en qué medida un gene extranjero $G_{a,i}$ cualquiera utiliza cada uno de los 64 codones del genoma receptor *b*. Definimos entonces el CRI del gene $G_{a,i}$ estimado con base en las frecuencias de codones del genoma *b* como:

$$CRI_b(G_{a,i}) = \sum_{c=1}^{64} p_b(c) * q_{a,i}(c). \qquad (1.1)$$

El índice puede interpretarse como la utilidad esperada de una distribución particular de codones y constituye una función de ponderación local [102], donde valores más grandes se obtienen cuando los codones más abundantes en el genoma *b* son utilizados por el gene $G_{a,i}$. El índice refleja, entonces, como "ve" el genoma *b* la composición particular de codones de $G_{a,i}$. Esto lo definimos como el "Potencial de Transferencia" del gene $G_{a,i}$ al genoma *b*. Debido a que las frecuencias $p_b(c)$ tienden a ser muy pequeñas, en la práctica los cálculos se realizaron multiplicando por 10 cada valor $p_b(c)$. Naturalmente, si se desea estimar el CRI de $G_{a,i}$ con respecto a su propio genoma *a*, el cálculo se debe realizar evaluando la Ecuación 1.1 para *b*=*a*, es decir $CRI_a(G_{a,i})$. Por ejemplo, la Figura 1.3 ilustra CRI para todos los genes de *E. coli* K12, tomando como referencia las abundancias globales de codones en *E. coli*.

Si un gene tuviera un UC aleatorio uniforme, donde todos los codones se utilizaran en la misma proporción, tendría un CRI de 0.15 con respecto a *E. coli*; un valor muy bajo como se puede apreciar en la Figura 1.3. El CRI teórico más alto posible sería para un gene que utilizara exclusivamente un codón, el más abundante en el genoma. En *E. coli* este valor sería CRI=0.528 que corresponde a frecuencia multiplicada por 10 del codón CUG para leucina. Obviamente, este es un valor teórico. Biológicamente, el CRI más elevado que se observa en *E. coli* es 0.2778 y el más bajo es 0.1464 (Figura 1.3).

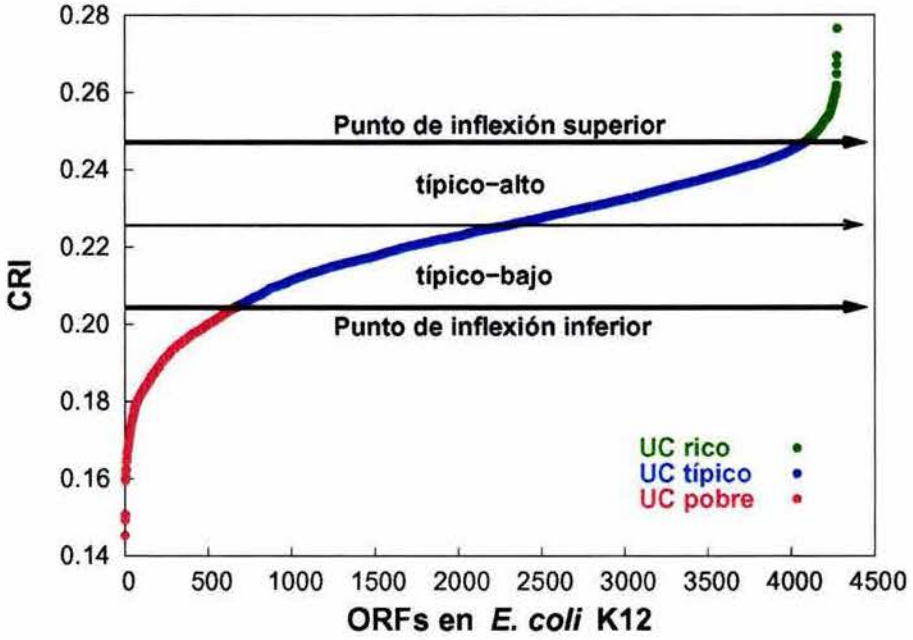

Figura 1.3. El índice de riqueza de codones (CRI) de los genes en *E. coli* K12. El eje de las Xs representa los ORFs ordenados ascendentemente de acuerdo a su valor de $CRI_a(G_{a,i})$ (eje de las Ys). Los puntos de inflexión de la curva son usados como umbrales para clasificar genes con alto, típico y bajo CRI (ver definición formal en la Sección 1.5.2). El punto medio entre los umbrales delimita las zonas de CRI típico–alto y típico–bajo.

Para descartar posible ruido en el cálculo del CRI debido a una anotación pobre de los genomas (i.e. genes falsos), inicialmente se tomaron en cuenta únicamente "genes reales", definidos como aquellos genes anotados con una longitud mínima de 600 pb y que tuvieran al menos 2 homólogos con una identidad menor al 80% para evitar efectos debidos a genomas redundantes. Se utilizó el programa BlastP [103] con un valor máximo de e-value igual a 10^{-3}, el alineamiento con cualquier candidato a homólogo debe cubrir al menos 50% de la proteína más pequeña. Sin embargo, al repetir el cálculo del CRI, considerando todos los genes anotados en cada genoma, se obtuvo una correlación superior a 0.99 entre los CRIs estimados por los dos enfoques. Por lo tanto, en los análisis aquí presentados se utilizaron todos los genes anotados en las cálculos.

1.5.2 Definición de los niveles pobre, típico y rico de UC.

Inicialmente se observó que si se ordenan todos los genes de *Escherichia coli* K12 de manera ascendente en base al índice de riqueza de codones (CRI), la mayoría de los genes despliega una diferencia aproximadamente constante de CRI (ver la zona de la curva en la Figura 1.3 donde la pendiente es lineal). Los genes con los valores de CRI más altos y más bajos exhiben diferencias más grandes de CRI y por lo tanto cambian significativamente la pendiente. Cuando se dibujaron curvas similares para todos los genomas fue posible observar, en casi todos los casos, que la zona intermedia entre los dos puntos de inflexión, donde la pendiente es aproximadamente constante, comprende aproximadamente el 80% del total de genes (ver Figuras 1.3 y 1.4). Esta observación sugiere como consecuencia una estrategia para la definición de umbrales, únicos a cada organismo, que permitan distinguir genes con alto y bajo CRI.

La Figura 1.4 muestra curvas del tipo observado en la Figura 1.3 para una muestra de 10 genomas incluyendo gram–positivas, gram–negativas y arqueas. Para cada genoma se trazó el histograma de su CRI y se calculó el rango de máxima densidad y mínima longitud que contuviera el 80% de los genes (ver Figura 1.5). Empíricamente se determinó que manejar histogramas de 40 intervalos permite obtener una mejor resolución, curvas más uniformes y minimizar brincos bruscos. Los intervalos fueron ordenados de mayor a menor de acuerdo al número de genes contenido en cada uno de ellos, de igual forma, los genes dentro de cada intervalo también fueron ordenados descendentemente. Posteriormente, partiendo del intervalo más grande, se fueron acumulando genes progresivamente, de acuerdo al ordenamiento de los intervalos, hasta obtener el 80% del total de genes, y finalmente se obtuvieron los valores de CRI mínimo y máximo de los genes seleccionados para definir los umbrales de CRI bajo y alto que se observan en la Figuras 1.3 y 1.4. Este procedimiento encuentra el conjunto de genes mostrando mínima variación en CRI, que corresponde a lo que se definió como la región de CRI típico. El punto medio entre los dos umbrales obtenidos se utilizó como frontera entre las zonas del CRI típico–bajo y típico–alto (ver Figuras 1.3 y 1.4).

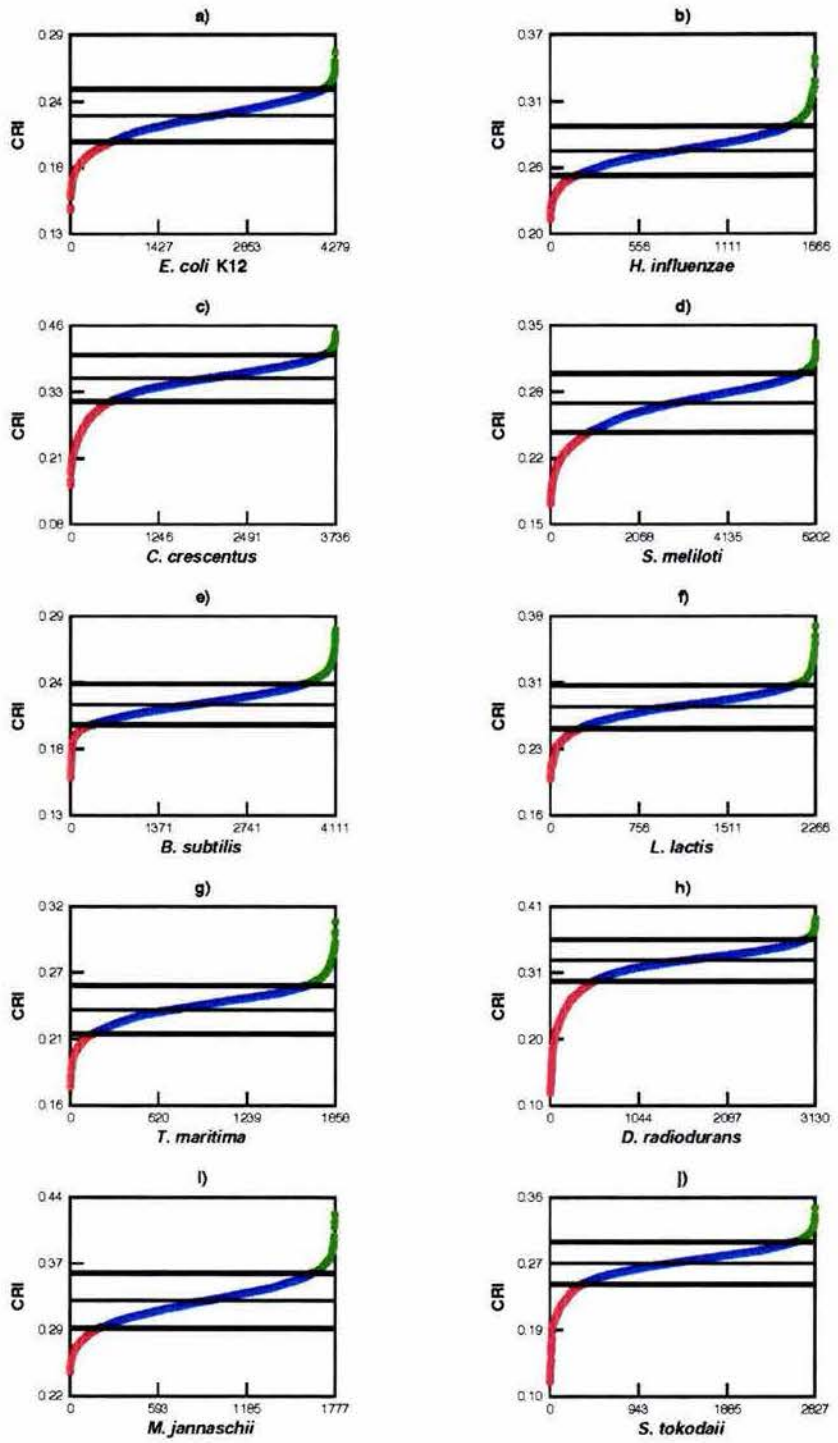

Figura 1.4. Los umbrales de alto y bajo CRI (líneas gruesas) abarcan el 80% de los genes en cada genoma. Independientemente de las tendencias particulares de cada distribución, los umbrales delimitan la zona lineal de cada curva (CRI típico). El código de colores es el mismo que en la Figura 1.3. El eje de las Xs indica los ORFs de cada genoma ordenados ascendentemente de acuerdo a su propio CRI. El eje de las Ys muestra los valores $CRI_a(G_{a,i})$.

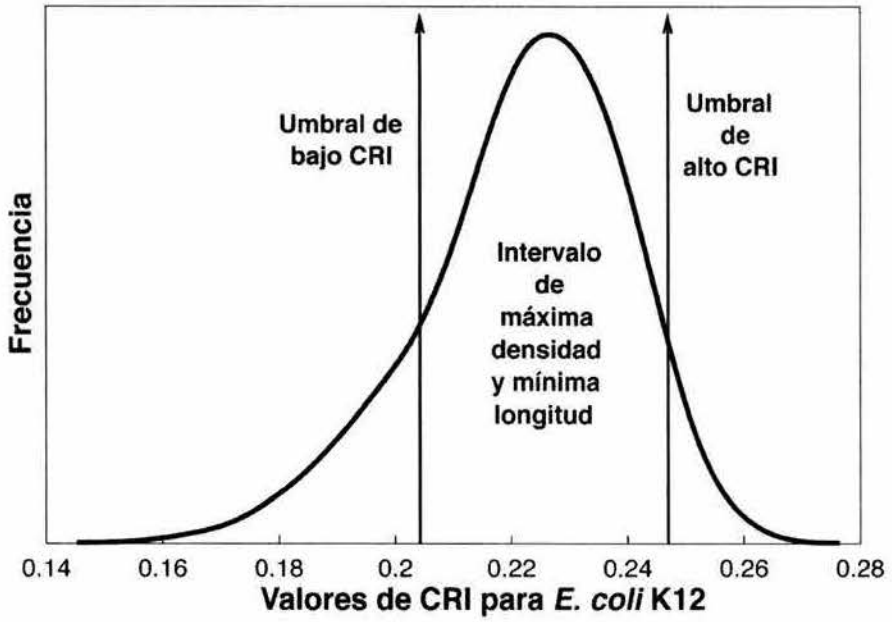

Figura 1.5. Histograma de frecuencias que ilustra como obtener los umbrales de bajo y alto CRI. Se calculan los puntos que delimitan el intervalo de máxima densidad y mínima longitud conteniendo el 80% de los genes. Los genes más desviados hacia valores más bajos de CRI serían predichos como foráneos por las metodologías actuales basadas en el paradigma composicional [34, 36].

En la Figura 1.4 el CRI de los distintos organismos no es comparable cuantitativamente. Al graficar para cada genoma los valores $CRI_a(G_{a,i})$ (ver Sección 1.5.1), se generaron escalas distintas para cada especie. A este nivel la comparación debe ser necesariamente cualitativa; por ejemplo, es informativo decir que dos genes ortólogos (ver Figura 1.1) tienen un CRI alto en sus respectivos genomas. Este problema se presentaría de cualquier forma si el CRI se normalizara para tener una escala entre 0 y 1, ningún índice de uso de codones quedaría exento de esta restricción. No sería informativo decir que dos genes ortólogos tienen un CAI (ver Sección 2.6.2) de 0.6 en sus respectivas especies porque el umbral para decir que un gene tiene alto CAI es diferente para cada organismo. Para que el CRI se pueda comparar numéricamente entre dos genes de distintas especies, es necesario calcular los valores $CRI_b(G_{a,i})$ como se muestra en las Secciones 1.6 y 1.8. Los cambios de escala en la Figura 1.4. sólo reflejan distintos sesgos en uso de codones. Por ejemplo, el codón más frecuente en *E. coli* (panel a), CUG, tiene una frecuencia (multiplicada por 10) de 0.5280, mientras que el codón más frecuente en *M. jannaschii* (panel i), AAA, tiene una frecuencia de 0.7261.

Tomando como referencia a *E. coli*, entre los genes con alto CRI se encuentran algunas proteínas ribosomales, factores de traducción involucrados en la iniciación, elongación y terminación, chaperoninas, tRNA sintetasas, proteínas de membrana, enzimas del ciclo de Krebs, glicólisis y otras vías metabólicas, subunidades α y β' de la RNA polimerasa, factores σ^{70} y σ^{32}, DNA y RNA helicasas, girasas, genes involucrados en división celular, algunos transportadores, proteínas de choque térmico, reguladores transcripcionales, algunas proteínas relacionadas con fagos, proteínas de flagelo y genes hipotéticos. Un 14% de los genes tienen función desconocida y otro 14% función putativa. Entre los genes con bajo CRI, hay 3 proteínas ribosomales (dos son putativas), algunas proteínas involucradas con la reparación, degradación y replicación del DNA, división celular, proteínas de membrana, enzimas varias, péptidos líderes, reguladores transcripcionales, proteínas de resistencia a antibióticos, transportadores, proteínas de fagos, transposones, muerte celular. Nótese que más del 50% de los genes con bajo CRI tienen función desconocida y el 25% tienen funciones putativas. La tendencia es clara, genes con alto CRI tienden a tener funciones más esenciales que genes con bajo CRI, pero esta fuera de los objetivos de esta tesis realizar un análisis más minucioso para determinar la generalidad de esta observación (ver Capítulo II, Sección 2.11).

Al utilizar en *E. coli* K12 los criterios de corte descritos en esta sección, los genes predichos como altamente expresados en base al sesgo de UC [30, 104] muestran un CRI típico–alto o rico (ver Capítulo II, Sección 2.6, Figura 2.3), mientras que los genes con un CRI típico–bajo o pobre correlacionan con genes predichos como adquiridos horizontalmente [34, 36]. Es decir, el 91% de los genes predichos como altamente expresados [104] tienen un CRI típico–alto o rico (para una discusión detallada ver Capitulo II, Secciones 2.6 y 2.8), mientras que un 84% de los genes predichos como adquiridos horizontalmente por metodologías composicionales [34] tienen un CRI típico–bajo o pobre. En la siguiente sección se dan más detalles de esta observación.

1.6 El potencial de Transferencia Horizontal

Un punto de referencia importante para interpretar cualquier tendencia en la composición de codones de genes que se mueven lateralmente, es conocer la probabilidad al azar de que un gene foráneo exhiba cualquiera de los tres niveles de UC al momento de ser adquirido. Como se mostró en las Secciones 1.5.1 y 1.5.2, los valores $CRI_b(G_{a,i})$ estimados mediante la Ecuación 1.1,

pueden usarse para determinar el nivel de UC que un gene foráneo $G_{a,i}$ desplegaría en cualquier genoma receptor b si sucediera un intercambio horizontal en este preciso instante. Es decir, al realizar una transferencia horizontal *in silico* del gene $G_{a,i}$ al genoma b se obtiene su potencial de transferencia.

Más formalmente, si se tienen n genomas, se considera a un genoma b, b=1, como referencia (receptor) y a todos los demás genomas a como donadores, donde $a=b+1,b+2...b+(n-1)$. La Ecuación 1.1 se aplica entonces para calcular todos los valores $CRI_b(G_{a,i})$, que posteriormente son clasificados en los tres niveles respectivos de UC. Este es el potencial de THG, en contraposición con eventos predichos de THG detallados en la Secciones 1.8 y 1.11. La Figura 1.6 muestra en 6 paneles cinco casos de distintos potenciales de transferencia horizontal. El caso menos frecuente (Figura 1.6, panel A) se ilustra con el donador hipotético *Ureaplasma urealyticum* cuyos genes en su mayoría pueden llegar directamente con un UC rico al genoma receptor, en este caso *Streptococcus pyogenes*. Este tipo de "afinidades" entre genomas constituyen solo el 4% de las potenciales transferencias cuando se toma en cuanta a todos los organismos (ver Tabla 1.1). La Figura 1.7 presenta las frecuencias de codones para estos dos genomas, sugiriendo al menos dos explicaciones para este comportamiento. Primero, *U. urealyticum* comparte todos los codones más abundantes con *S. pyogenes* y además los usa en frecuencias significativamente mayores (con la excepción de GCU). Segundo, los dos genomas también comparten los codones más raros, pero *U. urealyticum* los utiliza en frecuencias mucho menores que *S. pyogenes* (con la excepción UGA). El resto de los casos (Figura 1.6, paneles B-F) se pueden explicar por variaciones en estas dos condiciones extremas. Por ejemplo, el caso más frecuente, 74% del total (ver Tabla 1.1), se da cuando la mayoría de los genes de un potencial genoma donador llegan al genoma receptor con un UC muy pobre (Figura 1.6, paneles E-F). En esta situación los codones más abundantes de un genoma son los más raros del otro.

La Tabla 1.1, muestra el potencial de THG para todos los genomas. Claramente, desde la perspectiva del genoma receptor el 74% de los genes en otros genomas (potenciales donadores) presentan un UC pobre, el 22% es visto como genes con un UC típico y un 4% como genes con un UC rico. Si se restringe el cálculo del potencial de THG a genomas pertenecientes a la misma categoría taxonómica, por ejemplo solo entre proteobacterias, los resultados siguen un patrón similar, es decir que la mayoría de los xenólogos potenciales ($\geq$ 60%) exhiben un UC pobre en el

genoma receptor (ver Tabla 1.2); sin embargo, como es de esperarse, si los genomas son muy cercanos filogenéticamente el número de xenólogos potenciales con UC típico crece.

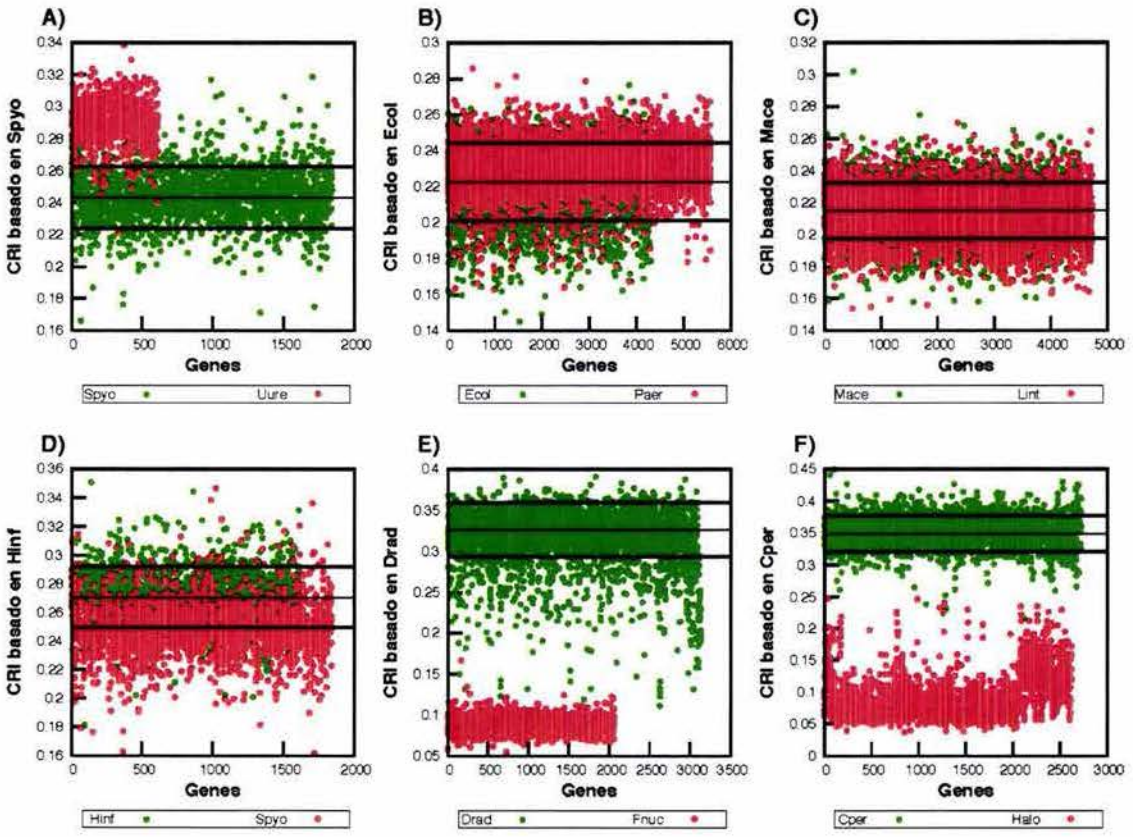

Figura 1.6. El potencial de transferencia horizontal. En el eje de las Ys se grafican como puntos rojos los valores $CRI_b(G_{a,i})$ para cada gene $G_{a,i}$ en el genoma donador *a* y como puntos verdes los valores $CRI_b(G_{b,i})$ para cada gene $G_{b,i}$ en el genoma receptor *b*. El eje de las Xs muestra el total de genes anotado en los genomas *a* y *b* en el respectivo orden cromosomal. Nótese que los genomas pueden tener muy distinto número de genes (Paneles A y E) y que cuando los genomas son razonablemente compatibles (Paneles B, C y D) los genes foráneos (puntos rojos) cubren visualmente a los genes nativos (puntos verdes). Los casos como el que se presenta en el Panel A son solo el 4% del total; casos como en los Paneles B, C, y D constitutyen un 22% del total; Los Paneles E y F representan el 74 del total de potenciales transferencias (Ver Tabla 1.1). Las líneas horizontales gruesas denotan los umbrales de alto y bajo CRI para el genoma receptor (ver Sección 1.5.2). Los genomas comparados son: *S. pyogenes* MGAS8232 (Spyo), *U. urealyticum* (Uure), *E. coli* K12 (Ecol), *P. aeruginosa* (Paer), *M. acetivorans* (Mace), *L. interrogans* (Lint), *H. influenzae* (Hinf.), *S. pyogenes* MGAS88232 (Spyo), *D. radiodurans* (Drad), *C. perfringes* (Cper) y *Holobacterum sp* (Halo).

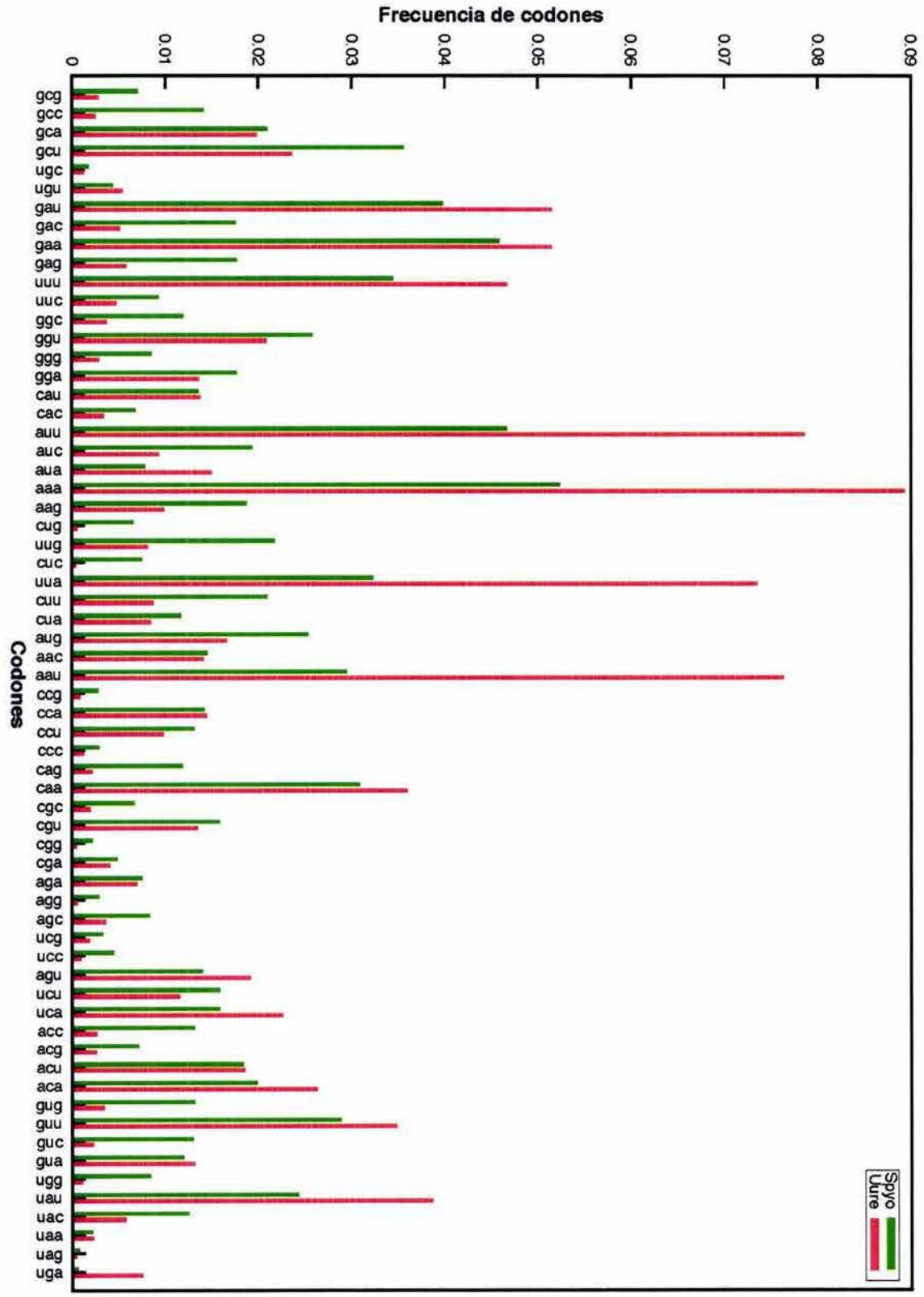

Figura 1.7. Frecuencias de codones de *S. pyogenes* MGAS8232 (Spyo) y *U. urealyticum* (Uure).

Tabla 1.1. El potencial de transferencia horizontal. Extracto de la Tabla 1 en [1] para 103 genomas procarióticos no redundantes (ver Sección 1.9.1.2). Se muestra que el número de genes foráneos que entrarían al genoma receptor (columna 1) con un CRI bajo (columna 2) es muy superior al número de genes que entrarían directamente con un CRI típico o alto (columnas 3 y 4 respectivamente). El total de genes foráneos que participaron en los cálculos se muestra en la columna 4 (excluyendo a los genes nativos del respectivo genoma receptor en turno).

Genoma	CRI bajo	CRI típico	CRI alto	Total
A. pernix	220691	73085	14	293790
S. solfataricus	232056	49256	11342	292654
Halobacterium sp	227533	65423	53	293009
M. thermoautotrophicum	282827	10809	122	293758
P. furiosus	231181	57219	5106	293506
T. acidophilum	279740	14265	144	294149
B. longum	192800	94234	6870	293904
C. glutamicum	157905	129792	4941	292638
S. coelicolor	262610	24695	118	287423
A. aeolicus	270363	23211	497	294071
C. pneumoniae TW 183	182254	66811	45453	294518
Synechocystis PCC6803	215477	73496	3491	292464
D. radiodurans	237132	51092	4276	292500
B. halodurans	174010	86401	31154	291565
B. subtilis	175368	96220	19931	291519
C. perfringens	273793	17441	1674	292908
L. innocua	226505	55510	10555	292570
M. genitalium	250284	39796	5067	295147
S. aureus Mu50	243684	43046	6153	292883
S. pyogenes MGAS8232	205537	69398	18851	293786
U. urealyticum	288095	6784	138	295017
A. tumefaciens C58 UWash	190657	72356	27216	290229
B. japonicum	206215	75167	5932	287314
B. melitensis	194500	71907	26029	292436
Buchnera sp	276040	17628	1389	295057
C. jejuni	265695	26368	1914	293977
C. crescentus	240000	51026	868	291894
E. coli K12	149254	135551	6515	291320
H. influenzae	236733	49700	7587	294020
H. pylori 26695	238771	49350	5946	294067
M. loti	201651	76917	9788	288356
N. meningitidis MC58	154865	138297	390	293552
N. europaea	150040	132992	10138	293170
P. aeruginosa	251160	36938	1966	290064
R. solanacearum	230872	58669	970	290511
R. conorii	234721	52704	6832	294257
S. meliloti	203487	66631	19310	289428
V. cholerae	133843	152096	5872	291811
W. succinogenes	286095	7412	80	293587
X. fastidiosa	77286	156537	58976	292799
Y. pestis CO92	102650	181804	6960	291414
T. maritima	263902	29201	670	293773
Promedio	216081.70	66000.35	10678.75	292760.80
Proporción	0.74	0.22	0.04	1.00

Tabla 1.2. El potencial de THG entre proteobacterias. Las columnas se leen igual que en la Tabla 1.1. De suceder una THG entre cualquier par de genomas en este linaje, existe una posibilidad del 66% de que el gene(s) involucrado(s) llegue(n) el genoma receptor mostrando un UC pobre (CRI bajo).

Genome	CRI bajo	CRI típico	CRI alto	Total
A. tumefaciens C58 UWash	54633	47435	22330	124398
B. floridanus	125835	2593	784	129212
B. bronchiseptica	99007	25541	247	124795
B. japonicum	64253	52931	4299	121483
B. melitensis	56189	47435	22981	126605
B. aphidicola	125986	2793	517	129296
Buchnera_sp	125664	3169	393	129226
C. jejuni	123847	3944	355	128146
C. crescentus	88116	37233	714	126063
C. violaceum	86603	38485	305	125393
C. burnetii	98058	23425	6308	127791
E. coli K12	34728	84496	6265	125489
G. sulfurreducens	58043	57281	11031	126355
H. ducreyi 35000HP	114419	12202	1462	128083
H. influenzae	116350	10116	1723	128189
H. hepaticus	118027	7480	2418	127925
H. pylori 26695	116150	10780	1306	128236
M. loti	61145	53737	7643	122525
N. meningitides MC58	39165	88220	336	127721
N. europaea	37976	79801	9562	127339
P. multocida	115324	10340	2121	127785
P. luminescens	88470	30519	6128	125117
P. aeruginosa	96106	26363	1764	124233
P. putida KT2440	65978	51642	6830	124450
P. syringae	51359	54007	18826	124192
R. solanacearum	81527	42239	914	124680
R. palustris CGA009	77841	40227	6915	124983
R. conorii	118405	8654	1367	128426
S. oneidensis	77858	41085	6388	125331
S. meliloti	62619	46819	14159	123597
V. cholerae	49129	74379	2472	125980
V. vulnificus YJ016	62752	58149	3871	124772
W. brevipalpis	128219	966	4	129189
W. succinogenes	124754	2988	14	127756
X. citri	78360	42996	4132	125488
X. fastidiosa	11196	69589	46183	126968
Y. pestis CO92	29456	92309	3818	125583
Promedio	29743.17	13421.05	2202.77	45366.99
Proporción	0.66	0.30	0.04	1.00

El potencial de THG refleja únicamente la probabilidad de que al tomar un gene al azar y transferirlo de un genoma a otro, este llegue con un uso de codones pobre, típico o rico. Obviamente, esta probabilidad no indica si la célula será capaz de expresar el gene, y mucho menos si éste confiere o no una ventaja selectiva al organismo receptor.

Como se describió previamente [1], el potencial de THG es consistente con los resultados obtenidos por varios autores quienes basaron sus predicciones de eventos recientes de THG en la ocurrencia de características atípicas en las secuencias de los genes [34, 36, 104]. Es necesario enfatizar, sin embargo, que estos datos sólo representan un perfil potencial contra el cual debe compararse un conjunto predicho de eventos de THG (ver Secciones 1.8, 1.9 y 1.11).

1.7 Identificación de Probables Ortólogos (PO).

Identificar una relación de xenología entre dos genes requiere que se descarte previamente la posibilidad de ortología y de paralogía. Dos genes son definidos como parálogos si divergieron a partir de un evento de duplicación genética, mientras que dos genes son considerados ortólogos si divergieron a partir de un evento de especiación (ver Figura 1.1). Como se mencionó en la Sección 1.4, es necesario detectar genes foráneos que aun conserven la huella composicional del DNA donador. Esta condición debe reflejarse en elevados niveles de identidad a nivel de proteína, como producto de la exigencia de un elevado parecido en la secuencia de nucleótidos. Por lo tanto, xenólogos con estas características deben pertenecer al conjunto de genes identificados mediante *definiciones de trabajo* para detectar ortología.

Se combinaron dos estrategias para detectar ortología. La primera involucra a los pares de genes que son más parecidos bi–direccionalmente (BDBH por sus siglas en ingles *Bi-Directional Best Hit* [105]), cuando cada gene se compara por separado contra todos los genes del otro genoma (ver Figura 1.8, Panel A, para una definición técnica). La segunda estrategia se aplicó en los casos donde no se encontró un BDBH en un genoma blanco. Aquí se tomó como probable ortólogo al gene con la calificación de BlastP más alta en el genoma blanco, siempre y cuando no existiera un gene con una mejor calificación de BlastP dentro del genoma referencia. Esta definición de trabajo para identificar ortología se bautizó como "ortólogo mejor que parálogo", o OHTP por sus siglas en inglés *Ortholog Higher than Paralog* [106] (ver Figura 1.8, Panel B, para una definición técnica). Todas las comparaciones se realizaron usando el programa BlastP [103] con un umbral máximo de e–value $\leq 10^{-3}$, filtrando regiones poco informativas y activando la opción que realiza el alineamiento final mediante el algoritmo Smith–Waterman [107].

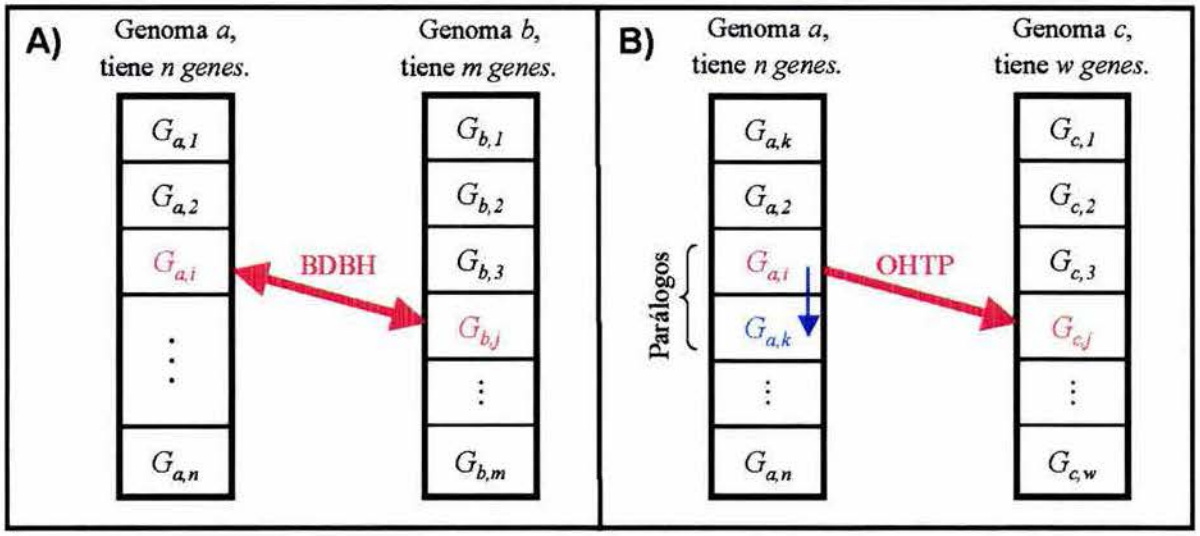

Figura 1.8. Identificación de Probables Ortólogos (PO) combinando dos definiciones de trabajo. A) Genes más parecidos bi–direccionalmente (BDBH por sus siglas en Inglés *Bi–Directional Best Hits*) [105]. Si el genoma *a* tiene *n* genes y *b* tiene *m*, el gene $G_{a,i}$ será BDBH de $G_{b,j}$ (genes en rojo) si al comparar $G_{a,i}$ contra todos los *m* genes en *b*, $G_{b,j}$ es el más parecido; y viceversa, cuando $G_{b,j}$ es comparado contra todos los *n* genes en *a*, $G_{a,i}$ es el más parecido. Esta relación se representa con la flecha roja de dos puntas. B) Ortólogo mejor que parálogo (OHTP por sus siglas en Inglés *Ortholog Higher Than Paralog*) [106]. Cuando no se encontró un BDBH para $G_{a,i}$ en un genoma *c*, se tomó como PO al gene más parecido en *c*, $G_{c,j}$ (ver flecha roja unidireccional vinculando a los dos genes rojos), siempre y cuando no existiera en el genoma *a* un gene $G_{a,k}$ (gene azul) con mayor parecido a $G_{a,i}$ que $G_{c,j}$ (flecha azul relacionando a dos genes dentro del genoma *a*). En este caso $G_{a,i}$ es OHTP de $G_{c,j}$. El grueso de las flechas roja y azul es proporcional al parecido entre los genes respectivos. Para interpretar las fórmulas descritas en la base de cada panel ver la discusión en el texto.

La relación de probable ortología aquí descrita, se puede formalizar para dos genes, $G_{a,i}$ y $G_{b,j}$, de la siguiente manera:

$$PO(G_{a,i},b) = \begin{cases} G_{b,j} \\ \phi \text{ (nulo de otra manera)} \end{cases}. \qquad (1.2)$$

Donde $PO(G_{a,i},b)$ es una función discreta que regresa, cuando existe, el probable ortólogo del gene $G_{a,i}$ en el genoma *b* ($G_{b,j}$) o un valor nulo de otra manera. La Figura 1.8 ilustra gráficamente esta relación de probable ortología.

1.8 Genes xenólogos recientes muestran un UC similar.

Con el propósito de averiguar si efectivamente un UC similar es una propiedad de genes xenólogos que aun conservan su UC original, se seleccionaron por cada par de genomas todos aquellos genes que son POs. Si dos genes son realmente xenólogos recientes, es claro que su similitud a nivel de DNA y proteína será tal que su relación puede ser confundida técnicamente por ortología. Por lo tanto, el conjunto de POs detectados con la metodología descrita en la Sección 1.7 es en esencia una mezcla de dos poblaciones: ortólogos y xenólogos.

Una vez identificados todos los pares de POs por cada par de genomas *a* y *b* ($G_{a,i}$, $PO(G_{a,i},b)$), se aplicó la Ecuación 1.1 para calcular el CRI de cada par de POs usando como referencia las frecuencias globales de codones en el genoma *a*; esto es, se obtuvieron los pares $[CRI_a(G_{a,i}),CRI_a(PO(G_{a,i}, b)]$. Posteriormente se calcularon las diferencias de CRI por cada par de POs usando la ecuación:

$$D_{a,i}(b) = CRI_a(G_{a,i}) - CRI_a(PO(G_{a,i},b)), \text{ para } PO(G_{a,i},b) \neq \emptyset. \qquad (1.3)$$

$D_{a,i}(b)$ es una función que regresa la diferencia del CRI, con base en el genoma *a*, entre el gene $G_{a,i}$ y su probable ortólogo en el genoma *b*, $PO(G_{a,i},b)$. Luego, se generaron histogramas con las diferencias absolutas de CRI $|D_{a,i}(b)|$, con el fin de observar tendencias en la similitud de UC. En la Figura 1.9 se muestra el histograma con las diferencias en CRI de todos los POs entre *Haemophilus influenzae* y *Neisseria meningitidis*, tomando como referencia las frecuencias genómicas de codones en *H. influenzae* (=*a*). Se pueden apreciar dos poblaciones en una distribución bimodal, donde una de ellas muestra diferencias cercanas a cero (línea punteada) y la otra población involucra diferencias considerablemente mayores (línea continua). De acuerdo con esta figura, los genes con $D_{a,i}(b)\approx 0$ serían candidatos as ser transferencias horizontales donde todavía se conserva la huella composicional del DNA donador. El histograma de la Figura 1.9 no es el único tipo de distribución encontrado, muchos pares de genomas no tienen POs con diferencias cercanas a cero (Figura 1.10, paneles A–B), sugiriendo que estos genomas no han intercambiado genes recientemente. Otros genomas tienen tan pocos POs con diferencias de CRI cercanas a cero, $D_{a,i}(b)\approx 0$, que la distribución no es bimodal (Figura 1.10, paneles C–D), indicando posiblemente la existencia de muy pocas transferencias o señales falsas. Algunos genomas tienen distribuciones bimodales similares a los de la Figura 1.9 (ver también Figura 1.10, panel E). Esta es la distribución que se esperaría encontrar en genomas que han

intercambiado clusters de genes, como operones o islas de patogenicidad. Por último, los POs de algunos genomas suficientemente cercanos muestran un número elevado de genes con diferencias de CRI muy pequeñas (Figura 1.10, panel F), sugiriendo que en estos casos un UC similar puede ser más el resultado de compartir un ancestro común reciente que de transferencia horizontal.

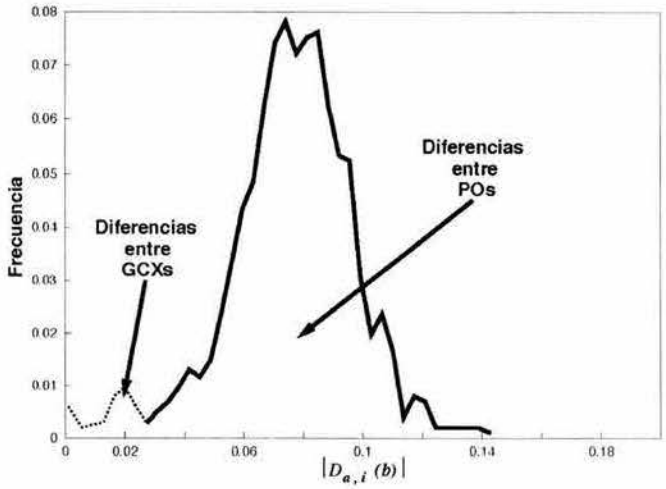

Figura 1.9. Diferencias absolutas de CRI para todos los pares de probables ortólogos (PO) entre el genoma referencia *a* (*H. influenzae*) y el potencial genoma donador *b* (*N. meningitidis*). Se propone que las diferencias más pequeñas (línea punteada) corresponden a genes candidatos a ser xenólogos (GCXs), los cuales todavía no han cambiado significativamente su composición original de codones, mientras que las diferencias más grandes (línea gruesa continua) corresponden a genes ortólogos, o bien a xenólogos ancestrales que ya han tenido mucho tiempo para divergir. Existen otros tipos de distribuciones, algunos ejemplos se dan en la Figura 1.10 y se discuten en el texto.

Histogramas como el de la Figura 1.10, panel F, indican claramente que el UC no es un parámetro discriminatorio confiable si se aplica en la ausencia de otros criterios o evidencias adicionales, ya que puede producir un número importante de falsos positivos —dos genomas pueden tener diferencias $D_{a,i}(b)$ muy pequeñas porque son parientes muy cercanos o simplemente porque su composición de G+C es tan similar que una diferencia de CRI cercana a cero puede ser producto del azar. Sin embargo, es seguro que de existir una transferencia reciente entre dos genomas, los genes intercambiados tendrán una diferencia $D_{a,i}(b)$ muy pequeña, independientemente de si las frecuencias genómicas de codones entre los organismos involucrados se parecen o no. Siguiendo esta línea de razonamiento, es posible asumir que entre más pequeña sea la diferencia de CRI entre dos POs, mayor será la posibilidad de que se trate de xenólogos donde todavía se conserva la composición de codones original del gene donador; por

supuesto los genes que cumplan esta condición deben ser sometidos a otros criterios adicionales, incluyendo el análisis filogenético, para obtener predicciones más confiables.

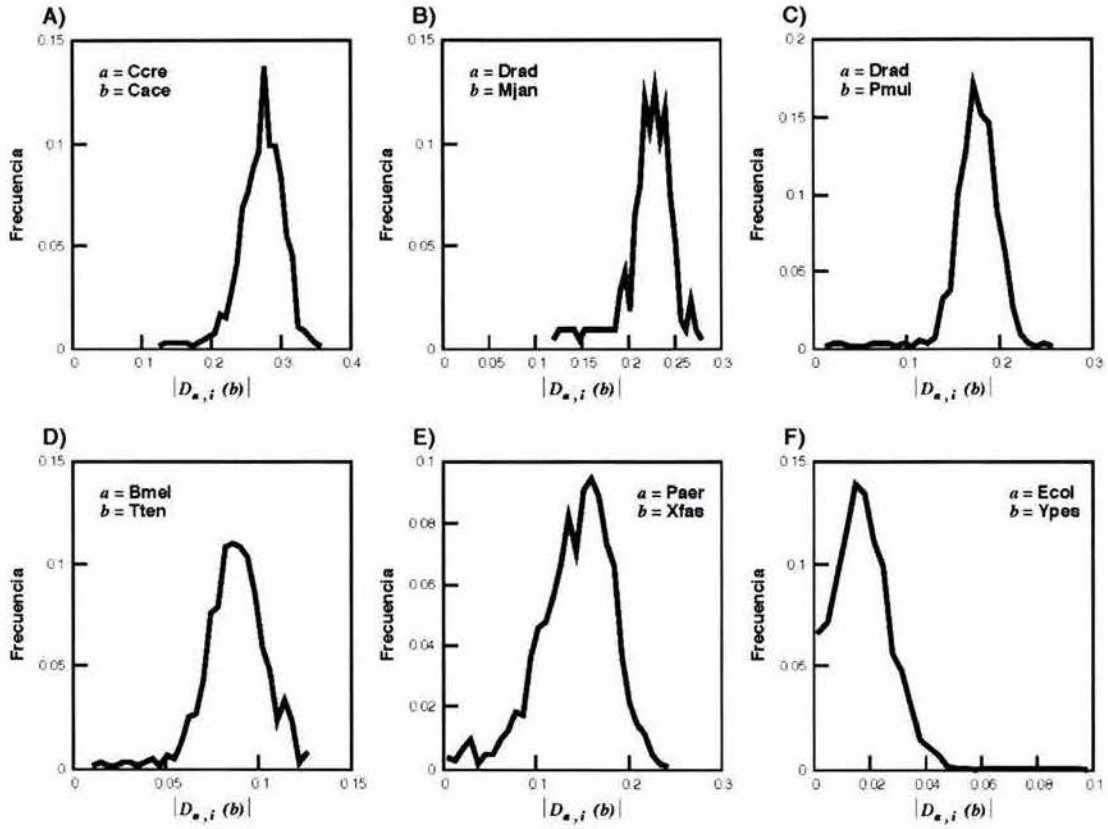

Figura 1.10. Histogramas mostrando las diferencias absolutas de CRI ($|D_{a,i}(b)|$) de todos los probables ortólogos (POs) entre pares de genomas. El genoma *a* es siempre el genoma referencia, cuyas frecuencias genómicas de codones se utilizan para estimar los valores $CRI_a(G_{a,i})$ y $CRI_a(\mathrm{PO}(\mathrm{G}_{a,i},\mathrm{b}))$. El genoma *b* es el genoma a comparar. Los organismos analizados son: *C. crescentus* (Ccre), *C. acetobutylicum* (Cace), *D. radiodurans* (Drad), *M. jannaschii* (Mjan), *P. multocida* (Pmul), *B. melitensis* (Bmel), *T. tengcongensis* (Tten), *P. aeruginosa* (Paer), *X. fastidiosa* (Xfas), *E. coli* K12 (Ecol) y *Y. pestis* CO92 (Ypes).

Finalmente, es necesario enfatizar que si dos probables ortólogos [$G_{a,i}$, $PO(G_{a,i}, b)$] muestran una diferencia $D_{a,i}(b)$ cercana a cero, no implica necesariamente que el genoma *b* es el donador —aunque efectivamente una interpretación es que tengan el mismo CRI porque el genoma *a* reconoce al genoma *b* como donador de $PO(G_{a,i}, b)$, es igualmente probable argumentar que el genoma *a* esta reconociendo al $PO(G_{a,i}, b)$ como propio, en cuyo caso el genoma *a* sería el donador.

1.9 Predicción de xenólogos

Los genes candidatos a ser xenólogos (GCXs) que aún conservan la huella composicional del DNA donador deben satisfacer al menos 4 requisitos. Primero, deben mostrar un uso de codones (UC) similar, cuantificado en este caso por medio del CRI ($D_{a,i}$(b)≈0) (Sección 1.8). Segundo, el tamaño debe ser aproximadamente el mismo (±10%) (Sección 1.9.1); longitudes diferentes indicarían que los genes han tenido la oportunidad de sufrir inserciones y/o pérdidas de nucleótidos importantes en su secuencia. Tercero, como consecuencia natural del parecido a nivel de DNA, los GCXs deben ser también los más parecidos a nivel de proteína cuando se realizan alineamientos globales entre un gene referencia $G_{a,i}$ y todos sus n POs involucrados en la predicción de un evento de transferencia horizontal de genes (THG) (Sección 1.9.1.1). No se realizan alineamientos locales porque se desea garantizar que los alineamientos incluyan las secuencias completas de los POs. Cuarto, la relación filogenética entre GCXs debe contradecir la filogenia estándar de las especies representadas por todos los POs (Sección 1.9.1.2).

Tomando en cuenta que el número de genomas totalmente secuenciados a la fecha no es suficientemente representativo, como para garantizar que se pueda determinar la identidad precisa de aquellos organismos involucrados en THG, siempre que aquí se prediga que un gene es intercambiado entre un par de genomas, el genoma donador es muy probablemente un pariente cercano del genoma que en realidad participó en la THG. Además, el sesgo ecológico inherente en el árbol de la vida, debido al número y tipo de genomas actualmente secuenciados, provocará forzosamente que se subestime del número real de eventos de THG.

La estrategia que se siguió para identificar candidatos a xenólogos (ver artículo en el Anexo I), donde todavía se conserva la huella composicional del gene donador, hace uso del teorema de Bayes. En la Sección 1.11, al final de este capítulo, se dan los conceptos básicos relacionados con teorema de Bayes y luego se realiza la deducción matemática del modelo estadístico empleado para identificar GCXs, utilizando las Ecuaciones 1.1, 1.2 y 1.3 como punto de partida. Los filtros adicionales a los que se sometieron los GCXs, identificados con base a un uso de codones similar, se describen en la Sección 1.9.1.

1.9.1 Filtros adicionales aplicados

Sólo se tomaron en cuenta predicciones donde hubiera al menos 5 unidades taxonómicas operativas (probables ortólogos), incluidos los dos genes candidatos a ser xenólogos (GCXs). Este criterio obedece a la recomendación de construir filogenias con al menos 5 genes [59]. Para garantizar que los GCXs no hayan sufrido inserciones o deleciones considerables en su secuencia, se exigió que todos los POs tuvieran longitudes similares (±10%).

1.9.1.1 Máximo parecido global entre GCXs

Construir filogenias puede ser una tarea que consuma una cantidad abrumadora de tiempo, especialmente si se deben construir miles de árboles. Por este motivo fue necesario elaborar una estrategia que prefiltrara los datos, evitando así validar filogenéticamente predicciones con pocas probabilidades de ser transferencias reales. Trabajando a nivel de aminoácidos, el plipéptido codificado por el gene referencia $G_{a.i}$ fue alineado globalmente contra los polipéptidos de todos sus probables ortólogos (POs) aplicando el algoritmo Needleman-Wunsch [108], implementado en el paquete EMBOSS [109], y utilizando los parámetros predefinidos. Se conservaron sólo aquellas predicciones donde los GCXs fueran los más parecidos por al menos una diferencia del 8% de identidad, con respecto al siguiente PO más cercano.

La diferencia del 8% es un umbral arbitrario y obedece al siguiente razonamiento. El que dos POs guarden la mayor similitud de secuencia no garantiza que compartirán un nodo terminal al realizar una filogenia [110]. Como consecuencia, se realizó un análisis para determinar cual es la diferencia en identidad global mínima necesaria para incrementar la probabilidad de que los GCXs compartan un nodo terminal al construir un árbol filogenético con todos los POs relacionados. Se tomaron al azar 100 predicciones que cumplieron la condición de mayor identidad a nivel global y se alinearon con ClustalW [111], usando los parámetros predeterminados. Posteriormente se generaron árboles con base en el método de agrupamiento de vecinos (*neighbor-joining*), eliminando posiciones con indels y usando la corrección para tasas desiguales de sustitución. La calidad de las topologías se evaluó con un *bootstrap*[9] de 1000 muestras. Cuando los GCXs muestran una diferencia de identidad global en el orden de 8% con

[9] Tipo de análisis estadístico para probar la confiabilidad de ciertas ramas en un árbol evolutivo. El proceso de bootstrap consiste en re-muestrear los datos originales, para crear una serie de muestras del mismo tamaño que los datos originales. El valor bootstrap de un nodo es el porcentaje de veces que ese nodo esta presente en el conjunto de árboles construido a partir de las muestras generadas.

respecto al siguiente PO más parecido, en el 90% de los casos los GCXs comparten un nodo terminal en el árbol resultante. El conjunto de predicciones que pasaron este filtro fueron posteriormente sometidos a una validación filogenética rigurosa.

1.9.1.2 Validación filogenética

El artículo adjunto en el Anexo I [1] discute con detalle la metodología empleada para realizar las pruebas de incongruencia filogenética. Aquí solo se harán algunos comentarios específicos que por cuestión de espacio no fueron incluidos en la publicación.

Un requisito fundamental para determinar si la historia evolutiva de un gene ha seguido una trayectoria vertical u horizontal, es tener una topología referencia que describa de manera confiable la filogenia de las especies bajo estudio. Esto puede ser un problema porque los criterios para el delineamiento de especies bacterianas aún se encuentran en franco proceso de maduración [112], especialmente cuando se trata de organismos tan cercanos que no es fácil decidir si se trata de dos especies o de dos cepas de una misma especie. Las dificultades se encuentran desde el inicio, por ejemplo, simplemente porque árboles de genes y árboles de especies no son necesariamente lo mismo [113]. Afortunadamente están surgiendo nuevas estrategias que combinan la hibridación DNA–DNA, criterios ecológicos (distribución geográfica), análisis filogenéticos y genética de poblaciones para lograr una mayor resolución. Un buen ejemplo es la reciente discriminación de especies dentro del género *Bradyrhizobium* [46].

Si los problemas para clasificar especies se dan principalmente entre genomas muy relacionados, una solución al conflicto es trabajar con especies que no son genéticamente redundantes. El número de genomas actualmente secuenciados (más de 200) y la existencia de técnicas para identificar genomas redundantes, acreditan esta idea como natural. En breve, el método para identificar genomas redundantes consta de dos pasos. Primero, obtener la calificación promedio de BlastP [103] de todos los BDBH (ver Sección 1.7) de un genoma contra si mismo, es decir la *auto–calificación promedio* (e.g. *E. coli* K12 vs *E. coli* K12). Segundo, obtener la calificación promedio de BlastP de todos los BDBH entre un genoma referencia y cualquier otro (e.g. *E. coli* K12 vs *D. radiodurans*), es decir la *calificación promedio comparada*. Finalmente dos genomas son considerados redundantes si el cociente de la calificación promedio comparada y la auto–calificación promedio es mayor o igual a 0.8 [114].

En aras del rigor científico, se compararon topologías tomadas de tres fuentes diferentes y el consenso se tomó como topología referencia para hacer los análsis filogenéticos. La primera de ellas, el árbol taxonómico estándar, se utilizó tal cual está reportado en la base de datos [115]. La segunda topología se obtuvo al construir un dendograma tomando como medida de distancia la similitud promedio de todas las proteínas compartidas por cada par de genomas [114] —ésta y la primer topología fueron extraordinariamente similares. La tercera topología se obtuvo al realizar filogenias rigurosas a nivel de proteína para los genes que se han propuesto como buenos marcadores moleculares [116]. La estrategia para inferir filogenias, con base en el criterio de máxima verosimilitud, permitiendo tasas variables de sustitución y con un análisis de 100 réplicas de bootstrap, está descrita en el artículo adjunto [1] en el Anexo I.

Muchos de los GCXs detectados generaron filogenias estrella (árboles donde se observan más agrupaciones "inesperadas" entre otros genes, sin contar al par de GCXs bajo análisis, y donde las ramas más profundas tienen poco soporte de bootstrap), este hecho ya ha sido considerado previamente como evidencia de transferencias laterales relativamente recientes [38]. Dada está situación, se decidió no aplicar pruebas estadísticas que comparen la topología referencia con la topología obtenida a partir de una predicción, pues la contribución a la incongruencia topológica no sería hecha únicamente por el par de GCXs predichos, sino por otros genes en el árbol. La topología referencia se utilizó para encontrar al menos un gene cuya especie esté más relacionada con alguno de los dos GCXs involucrados en una predicción de transferencia horizontal; se consideró a este criterio suficiente para sustentar la incongruencia filogenética de una predicción, siempre y cuando el soporte de bootstrap entre los dos GCXs fuera superior al 75%.

La Figura 1.11 muestra el caso donde un gene hipotético (gi:18977095) se transfiere entre la eubacteria *T. marítima* y el arquea *P. furiosus* (nombres resaltados en fondo amarillo), con un soporte de bootstrap muy elevado (96%). La relación incongruente se da porque *P. furiosus* debió haber quedado agrupado con el grupo de las arqueas y *T. marítima* con las bacterias. Esta predicción es confiable porque el grupo de las arqueas quedo óptimamente resuelto (bootstrap = 100%). Nótese que la agrupación mostrada por las bacterias *S. coelicolor* y *S. melitensis* con el linaje que lleva a las arqueas no es significativa (bootstrap = 14%).

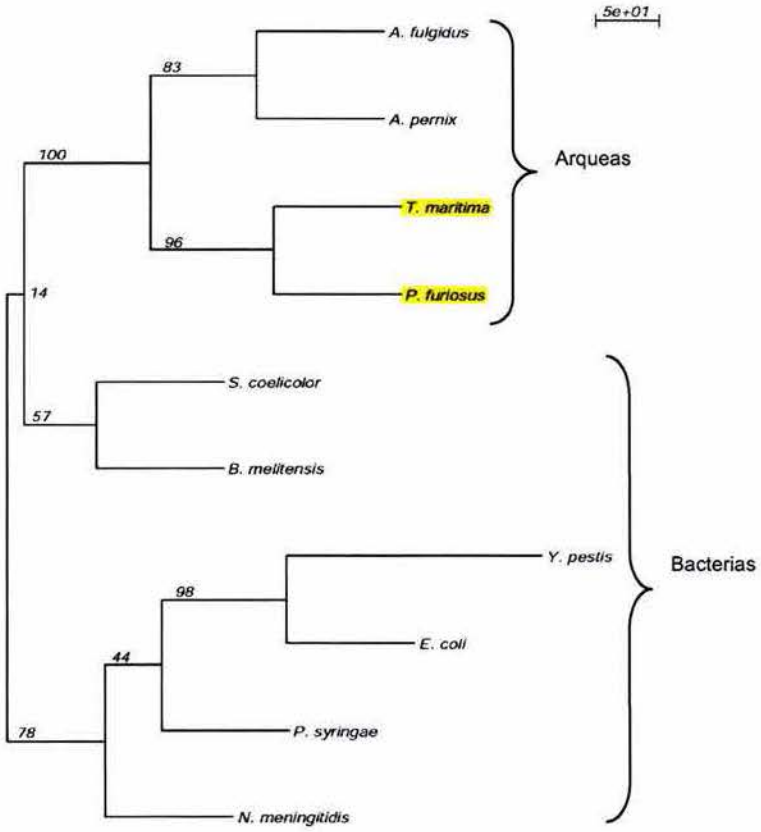

Figura 1.11. Transferencia del gene hipotético (gi:18977095) entre *P. furiosus*. Y *T. Maritima* (nombres resaltados en fondo amarillo). Los números (porcentajes) en los nodos y longitud de las ramas denotan los valores de bootstrap. Estos GCXs muestran una identidad global a nivel de proteína del 66.3% y su relación esta soportada por un bootstrap del96%. La discusión sobre la incongruencia filogenética se encuentra en el texto.

Otro ejemplo lo representa el intercambio horizontal de un regulador de la transcripción, perteneciente a la familia MarR/EmrR (gi:15895751), entre los genomas *C. acetobutylicum* y *L. lactis* (resaltados en fondo amarillo en la Figura 1.12). Este gene no se encuentra en el linaje *Clostridia* (no ilustrado en el árbol), excepto por *C. acetobutylicum*, mientras que sí está distribuido en el linaje *Bacilli*. Nótese que los linajes individuales de las bacterias gram–positivas y gram–negativas están resueltos razonablemente bien, con la excepción de *A. tumefaciens* (bootstrap = 16%) y *B. Anthracis* (bootstrap = 32%). A primera vista parece atractiva la idea de sugerir que *L. lactis* es el donador. Otro aspecto de este árbol es la relación incongruente entre *B. melitensis* y *S. meliloti* (marco rojo), porque en el árbol de la vida *A. tumefaciens* es más cercano a *S. meliloti*.

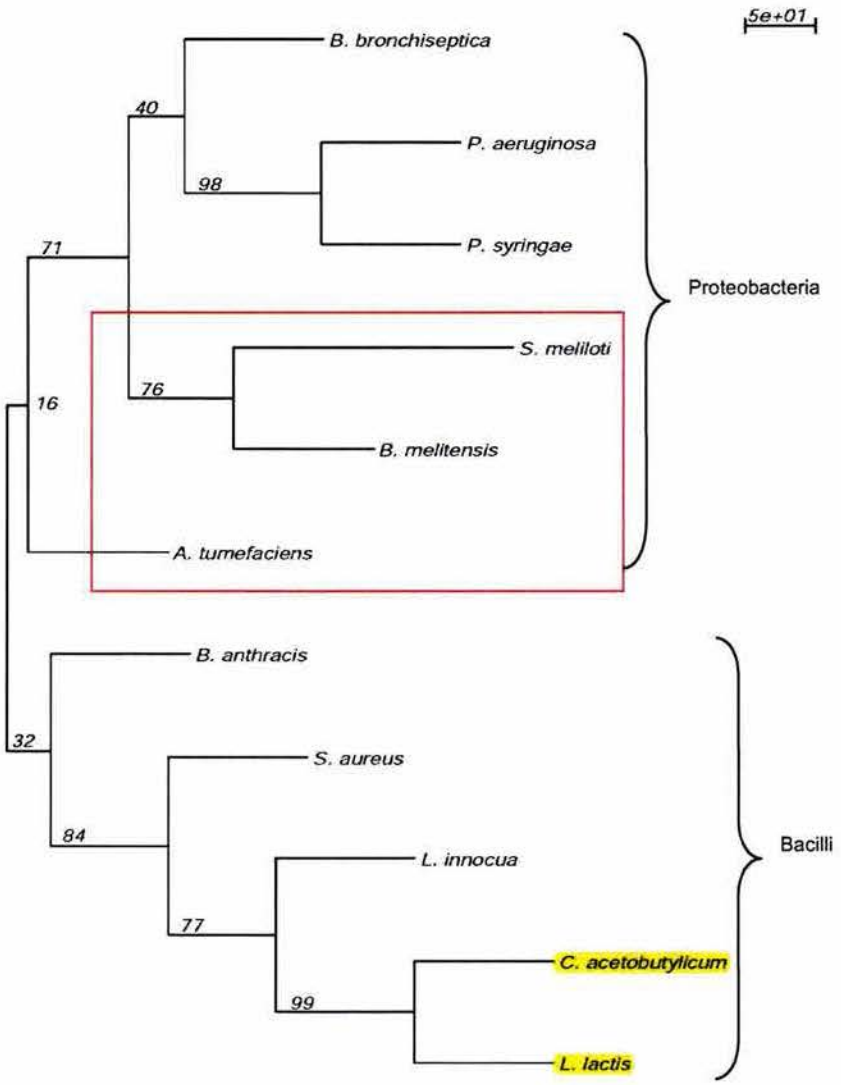

Figura 1.12. Trasferencia del gene regulador de la familia MarR/EmrR (gi:15895751) entre *C. acetobutylicum* y *L. lactis*. (nombres resaltados en fondo amarillo). Los dos genes guardan una identidad global a nivel de proteína del 50%. Los números (porcentajes) en los nodos y las longitudes de las ramas denotan el soporte de bootstrap. Se observa otra relación filogenética incongruente entre *B. melitensis* y *S. meliloti* (marco rojo). Ver discusión en el texto.

Un último ejemplo se ilustra en la Figura 1.13 con la transferencia de una acetiltransferasa (gi:24371794) entre las γ–proteobacterias *S. oneidensis* y *V. vulnificus* YJ016 (resaltados en fondo amarillo). El gene en *S. oneidensis* quedó en medio de las dos bacterias *Vibrio* (bootstrap=100%) estableciendo la anomalía filogenética. El porcentaje de identidad global entre los dos GCXs es 65% mientras que la identidad con el PO en *V. cholera* es del 49%. La rama más profunda del linaje de las proteobacterias no quedó bien resuelta (bootstrap=24%).

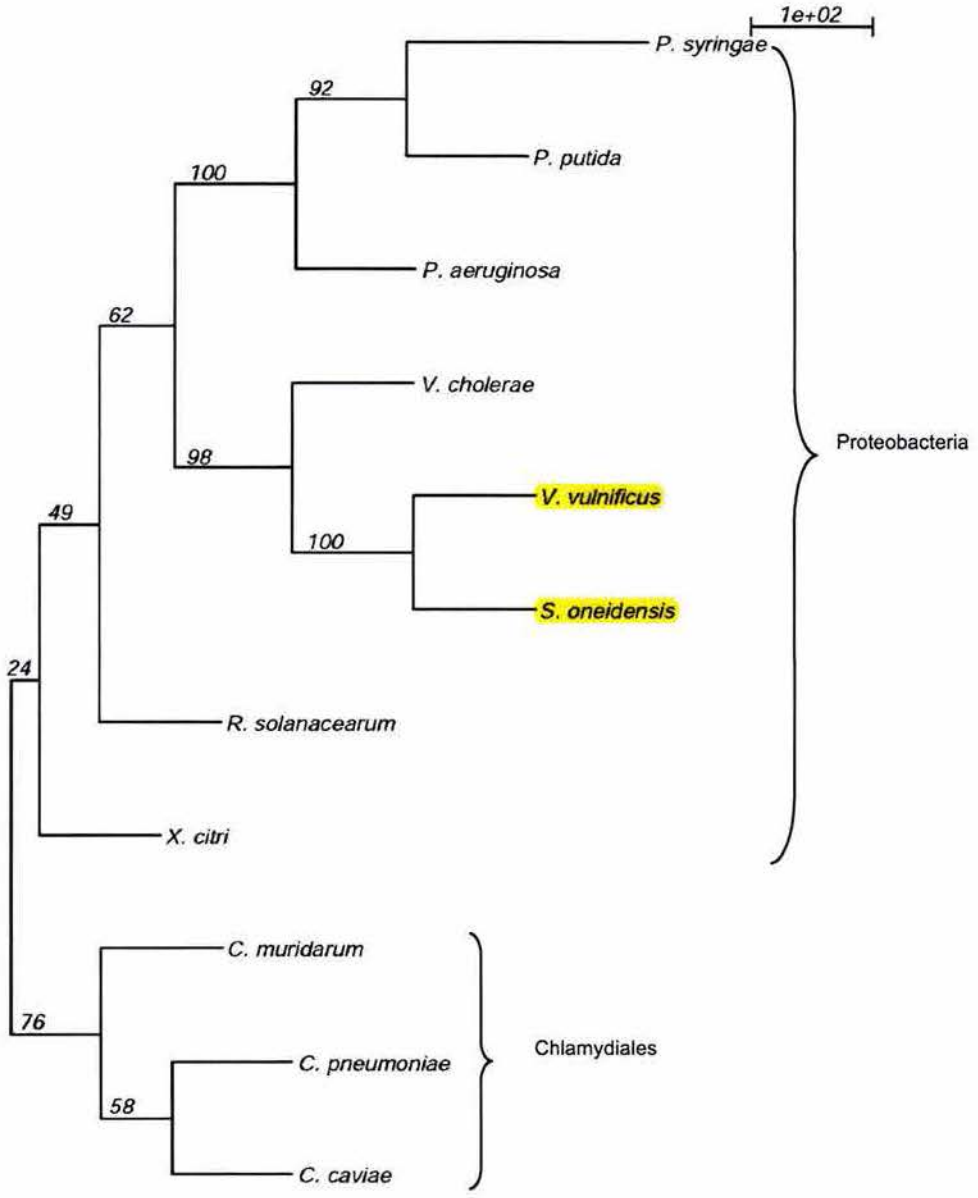

Figura 1.13. Transferencia de una acetiltransferasa, gi:24371794, entre *S. oneidensis*. Y *V. vulnificus* YJ016. Los dos genes guardan una identidad global a nivel de proteína del 65.5%. Los números (porcentajes) en los nodos y la longitud de las ramas reflejan el soporte de bootstrap. La relación de incongruencia está discutida en el texto.

El total de transferencias que pasan la prueba de incongruencia filogenética se muestra en la Tabla 1.3, la cual es un extracto de la Tabla 2 en el artículo incluido en el Anexo I [1]. Se puede apreciar claramente como el 89% de los genes muestran un UC típico o rico, contrastando notoriamente con el potencial de THG (Tabla 1.1) que estima una probabilidad de 0.74 de que los genes xenólogos muestren un UC pobre en el genoma receptor.

Tabla 1.3. Nivel de CRI de los genes predichos como involucrados en eventos de THGs mediante el criterio de similitud de UC (Secciones 1.4 a 1.9). La tabla es un extracto de la Tabla 2 en [1]. Para todos los genomas (columna 1), se muestra el número de genes involucrados en eventos de THG con un UC pobre (columna 2), típico (columna 3) y rico (columna 4). El total de genes involucrados en transferencias por genoma se muestra en la quinta columna.

Genoma	CRI bajo	CRI típico	CRI alto	Total
S. solfataricus	0	1	0	1
M. acetivorans	0	4	0	4
S. coelicolor	0	11	0	11
B. thetaiotaomicron VPI-5482	0	7	0	7
C. caviae	0	0	1	1
C. tepidum. TLS	0	8	1	9
G. violaceus	0	4	0	4
Nostoc. sp	0	2	1	3
D. radiodurans	2	1	0	3
B. anthracis A2012	0	5	0	5
B. halodurans	0	4	0	4
C. perfringens	0	9	0	9
E. faecalis V583	0	12	0	12
L. lactis	1	2	2	5
L. innocua	0	4	1	5
S. agalactiae 2603	0	6	1	7
S. pneumoniae R6	0	14	0	14
Pirellula. sp	0	6	1	7
A. tumefaciens C58 UWash	3	30	1	34
B. bronchiseptica	3	19	1	23
B. japonicum	2	29	1	32
B. melitensis	2	20	0	22
C. crescentus	0	10	0	10
C. violaceum	3	18	0	21
E. coli K12	1	20	0	21
G. sulfurreducens	1	8	0	9
H. influenzae	3	2	0	5
M. loti	4	57	2	63
N. meningitidis MC58	3	7	0	10
N. europaea	4	11	3	18
P. multocida	0	3	0	3
P. aeruginosa	3	22	3	28
P. putida KT2440	1	13	0	14
P. syringae	2	23	0	25
R. solanacearum	3	25	2	30
R. palustris CGA009	0	8	1	9
S. oneidensis	0	15	0	15
S. meliloti	3	49	7	59
V. vulnificus YJ016	1	8	1	10
X. citri	0	21	0	21
X. fastidiosa	0	2	2	4
Y. pestis CO92	0	13	0	13
T. maritima	0	5	1	6
Total	81	615	34	730
Fracción	0.11	0.84	0.05	1.0

1.10 Discusión

El resultado principal de esta tesis doctoral indica que la gran mayoría de los genes involucrados en eventos exitosos de transferencia horizontal, son genes que llegan al genoma receptor mostrando un UC principalmente típico [1]. Las tendencias fueron siempre las mismas independientemente del método de reconstrucción filogenética aplicado (i.e máxima parsimonia, agrupamiento de vecinos o máxima verosimilitud) y de la exigencia del nivel de identidad, a nivel de proteína, que se empleara para argumentar que una transferencia horizontal es reciente (ver Tabla 3 en el artículo adjunto en el Anexo I). El número total de genes involucrados en eventos de transferencia horizontal podría parecer muy pequeño (Tabla 1.3) considerando que el análisis se realizó usando 103 genomas no redundantes, pero esto se debe, entre otras razones, a los exigentes criterios que se usaron para filtrar las predicciones —(1) sólo se tomaron en cuenta genomas completos; (2) los grupos taxonómicos no están uniformemente representados en el conjunto de genomas analizados; (3) se requiere que cada par de candidatos a ser xenólogos comparta un nodo terminal en el árbol filogenético, (4) los genes deben ser de longitudes similares; y (5) debe haber al menos 5 probables ortólogos para realizar un análisis filogenético— y a que sólo aquellos pares de xenólogos donde aún se conserva la huella composicional del gene donador pueden servir para determinar si el UC de genes foráneos es típico o atípico inmediatamente después del evento de transferencia. Debido a que las transferencias detectadas involucran genes de todos los linajes procariotes representados en el conjunto de genomas analizados, se tiene confianza en que los resultados obtenidos son representativos y nos muestran la naturaleza composicional de la mayoría de los genes involucrados en intercambios horizontales exitosos.

Los resultados indican que los genes importados desde otras especies muestran predominantemente un UC típico en el genoma receptor en el instante de la transferencia. Esta afirmación desafía los supuestos fundamentales detrás del paradigma composicional (Sección 1.3), donde se enfatiza que, justo en el instante en que se da la transferencia, los genes foráneos despliegan un uso de codones pobremente adaptado al genoma receptor. De ser así, una implicación importante de los resultados, es que el nivel típico de UC parece definir una zona de seguridad o tolerancia donde los genes pueden alcanzar niveles adecuados de expresión. El Capítulo II, analiza más en detalle esta posibilidad al relacionar el UC con la concentración de tRNA y, entre otras cosas, se muestra que un UC típico efectivamente es suficiente para obtener

una eficiencia razonable de la traducción. Es decir, asumiendo que en el momento de la transferencia, de alguna manera la célula receptora resuelve los problemas genéticos respectivos para expresar los genes foráneos (e.g. promotor, regulador de la transcripción, sitio Shine–Dalgarno, etc.), el factor determinante para poder evaluar la función importada está ubicado en la etapa de traducción. Genes foráneos con una composición de codones razonablemente similar a la del genoma receptor podrían, en principio, ser más fácilmente expresados. En el artículo adjunto (Anexo I) se discuten con detalle las ventajas selectivas de los genes foráneos al ser compatibles con la maquinaria de traducción de la célula receptora.

Las funciones de los GCXs son congruentes con la hipótesis de complejidad [48] porque se trata principalmente de genes operacionales e hipotéticos (e.g. diversas enzimas metabólicas, transportadores, reguladores de la transcripción, genes de resistencia a antibióticos, elementos móviles, entre otros). En el artículo adjunto (Anexo I), se plantea que es poco probable que un gene foráneo llegue al genoma receptor con un UC pobre y que su secuencia evolucione tan rápidamente (para reflejar el UC promedio del nuevo contexto genómico) que la metodología reportada en este capítulo no pueda detectar este tipo de eventos. Aquí se mencionarán otros argumentos adicionales a favor de esta idea. En primer lugar, un estudio de 3,595 grupos de genes homólogos encontró que sólo el 0.5% de los genes (aquellos relacionados con el sistema inmune) muestran tasas rápidas de evolución [117]. En segundo lugar, aunque la discusión sobre si los genes esenciales evolucionan más lentamente que los no esenciales se puede rastrear a la década de los 70s [118], actualmente el tema sigue siendo muy controversial. Por ejemplo, mediante un estudio de genes en ratón, el grupo de Laurence Hurst [119] propuso que las tasas de evolución entre genes esenciales y no esenciales no son significativamente diferentes, posición que fue impugnada posteriormente por el grupo Aaron Hirsh [120], al afirmar que en levadura los genes más indispensables —aquellos que reducen la velocidad de crecimiento cuando son mutados— exhiben una tendencia gradual hacia variar más lentamente. Sin embargo, el grupo de Hurst repitió el análisis de Hirsh usando un conjunto más grande de datos y no encontró evidencia de que la dispensabilidad de proteínas pueda explicar la variación en tasas de evolución [121]. El debate continua. Por lo tanto, si las diferencias en tasas de evolución entre genes esenciales y no esenciales son insuficientes o muy pequeñas (ver respuesta de Hirsh [121]) como para apoyar contundentemente una u otra hipótesis, entonces, a menos de que exista una intensa presión selectiva para evolucionar muy rápido (e.g. antígenos de superficie en parásitos), o

ninguna presión en absoluto (e.g. pseudogenes), es poco probable que los genes foráneos lleguen al genoma receptor con UC pobre y varíen muy rápidamente para subir a la zona de UC típico.

La posibilidad de identificar pares de genes donador–receptor hace factible el estudio de redes de transferencia horizontal. Aunque tal estudio está fuera de los objetivos de esta tesis, no es difícil vislumbrar sus implicaciones. Conocer qué genomas intercambian qué tipo de genes con mayor o menor frecuencia, sería muy útil para rastrear que presiones selectivas han actuado sobre los genomas a lo largo de su evolución y para comprender mejor la dinámica del flujo horizontal de genes.

Si bien los resultados presentados en este capítulo entran en evidente conflicto con las metodologías composicionales, nuevas evidencias recientemente publicadas apoyan la noción de que la compatibilidad de UC entre genes foráneos y genomas receptores juega un papel importante. El potencial de THGs predice que la mayoría de las transferencias ocurren con genes que muestran composiciones atípicas en el genoma hospedero (Tabla 1.1 y 1.2), y los resultados finales indican que los pocos que logren llegar al genoma receptor con una composición de codones típica o rica tendrán más oportunidad de ser transferencias exitosas. Apoyando esta perspectiva, estudios comparativos recientes en bacterias han concluido que una cantidad considerable de los pseudogenes surgieron como producto de transferencias horizontales fallidas (están siendo eliminados por deriva génica) y que, en comparación con genes endógenos, tienen más del doble de probabilidad de mostrar un UC anómalo [122].

Como se mencionó en Antecedentes (Sección 1.3), actualmente el grueso de la literatura sugiere que la THG ha jugado un papel crítico en la divergencia de cepas bacterianas a partir de un ancestro común, pero sencillamente se desconoce si en efecto todos estos genes exógenos expresan de verdad proteínas funcionales. En un análisis genómico–funcional reciente se observó que sólo una minoría de los genes predichos como adquisiciones laterales en *E. coli* K12 (por criterios composicionales) muestran productos proteicos detectables, mientras que la gran mayoría muestra niveles de mRNA comparables con genes nativos. Los microarreglos y ensayos proteómicos se construyeron a partir de células cultivadas en un medio Luria–Bertani, tomando muestras en fase logarítmica tardía y fase estacionaria temprana. Los autores concluyen que la gran mayoría de los genes foráneos son incompatibles con la maquinaria de traducción del genoma receptor y por lo tanto no generan proteínas funcionales [123].

Un estudio de cómo han afectado los parámetros geográficos y ambientales la transferencia horizontal concluyó que la THG no es aleatoria, de hecho se encontró que depende críticamente de que exista una compatibilidad de factores internos y ambientales entre los organismos involucrados. Los más significativos son el tamaño del genoma, la composición de G+C, utilización de carbono y tolerancia al oxígeno. Al parecer, otros factores como la salinidad, temperatura, pH y distribución geográfica tienen solamente efectos débiles [124]. Estos resultados apoyan la noción de que debe haber ciertas compatibilidades entre genomas para incrementar las posibilidades de éxito de una transferencia horizontal. A su vez, la propuesta de compatibilidad de contenido de G+C por definición se contrapone a los métodos de predicción basados en el paradigma composicional (ver Sección 1.3.1.2) y apoya las conclusiones de este trabajo pues existe una correlación positiva no muy elevada pero estadísticamente significativa entre porcentaje de G+C y CRI ($r = 0.5$ en promedio para todos los genomas analizados; $p < 0.001$).

Una ventaja de los resultados aquí presentados es que pueden evaluarse fácilmente en el laboratorio. Por ejemplo, se pueden seleccionar dos genomas *a* y *b* con un uso de codones muy incompatible y un gene en copia sencilla que se exprese en cantidades considerables (e.g. una enzima de la vía de la glicólisis o del ciclo de Krebs). Después, se puede remplazar al gene en *a* por su ortólogo en *b* mediante mutagénesis sitio–dirigida sobre la misma región que contiene al gene en *a* (se podría probar también insertando al gene en otras regiones) —sería conveniente que el genoma *a* fuera *E. coli* dado que esta metodología experimental ya está bien probada en este organismo [125]. Otra alternativa es inactivar al gene en *a* y posteriormente introducir en *a* su ortólogo de *b* en un plásmido unicopia. El objetivo es ver si el gene en *a* logra complementar el fenotipo de su ortólogo en *b*. El experimento debe repetirse ahora para genomas con una composición de codones compatible y comparar los resultados. La predicción sería que los genes intercambiados entre genomas con uso de codones compatible podrán complementar el fenotipo mucho mejor que el intercambio entre genomas incompatibles.

Finalmente, es factible pensar en un proyecto para modelar la dinámica de la transferencia horizontal. Hay tres escenarios posibles: (1) el gene foráneo puede coexistir con el gene ortólogo en el genoma hospedero; (2) el gene foráneo puede desplazar al gene ortólogo; y (3) previamente no existía en el genoma receptor un gene homólogo al gene adquirido. La probabilidad de que exista la transferencia estaría en función de la posibilidad de contacto entre los organismos y de

su habilidad para adquirir DNA de otras especies mediante transformación, conjugación o transducción. La probabilidad de éxito de la transferencia sería una función que ponderara la ventaja funcional aportada por el gene adquirido, la magnitud de la presión selectiva ejercida por el medio ambiente, la tasa de mutación, la compatibilidad de uso de codones y de otras características previamente discutidas [124], entre otros factores. Se podría simular, con el paso del tiempo, la propagación del gene adquirido en la población del genoma receptor mediante muestreos a ambas funciones, pero el reto es en realidad determinar el tipo de relación matemática entre los diferentes parámetros.

1.11 Deducción Matemática del modelo Bayesiano empleado para identificar GCXs

1.11.1 Introducción al teorema de Bayes

Sea x_i un dato tomado al azar distribuido conforme a $F(\theta)$ —el parámetro θ no es directamente observable a partir de x_i. Ahora, si tomamos al azar una muestra $X=(x_1,x_2,..,x_n)$, de tamaño n, con la distribución $F(\theta)$ y asumimos que el parámetro θ es discreto pudiendo tomar únicamente un número v de valores, se puede plantear la hipótesis de que el parámetro θ tenga el valor específico b, $\theta=b$, y calcular su probabilidad, dados los datos en la muestra X, esto es $P(\theta=b \mid X)$. El teorema de Bayes permite realizar inferencias de este tipo sobre el parámetro poblacional θ, dada una hipótesis y una muestra X, mediante la aplicación de la siguiente ecuación:

$$P(\theta = b \mid X) = \frac{P(X \mid \theta = b)P(\theta = b)}{\sum_{j=1}^{v} P(X \mid \theta = j)P(\theta = j)}. \qquad (1.4)$$

Donde $P(\theta=b \mid X)$ se define como la probabilidad posterior de la hipótesis dada la muestra X. Es decir, la probabilidad de que el parámetro poblacional θ sea igual a b dados los datos x_i en la muestra X. El denominador de la Ecuación 1.4 es una constante de proporcionalidad y garantiza que $\sum_{j=1}^{v} P(\theta = j \mid X) = 1$, es decir, equivale a evaluar el numerador de la Ecuación 1.4 para cada uno de los v posibles valores de θ:

$$\omega = \sum_{j=1}^{v} P(X \mid \theta = j)P(\theta = j). \qquad (1.5)$$

$P(X \mid \theta=b)$ puede interpretarse como el modelo que describe el comportamiento estadístico de los datos dado θ —por ejemplo, si tenemos razones para suponer que la población F se comporta de manera Poisson, entonces el modelo $P(X \mid \theta)$ es la distribución Poisson y el parámetro θ correspondería a la media, λ. Asumiendo independencia estadística entre los datos de la muestra, tendríamos que:

$$P(X \mid \theta) = \prod_{i=1}^{n} P(x_i \mid \theta). \qquad (1.6)$$

$P(\theta=b)$ es la probabilidad *a priori* de θ, y puede interpretarse como el conocimiento previo, independiente a X, que tengamos sobre la probabilidad de que $\theta=b$.

1.11.2 Selección de POs con UC similar

Las Figuras 1.9 y 1.10, panel E, muestran que en general los genes xenólogos que aún conservan el rastro de su composición de codones original despliegan diferencias de CRI más pequeñas que los genes ortólogos. Entonces, tomando como referencia al gene $G_{a,i}$ se puede poner a competir a todos sus POs en otros genomas, de tal manera que aquel $PO(G_{a,i},b)$ que muestre la diferencia $D_{a,i}(b)$ más pequeña tendrá mayores oportunidades de ser el verdadero xenólogo. Esta competencia se modeló aplicando el teorema de Bayes con el fin de calcular la probabilidad de que dos genes sean xenólogos dadas las diferencias en CRI entre un gene referencia $G_{a,i}$ y todos sus n posibles POs en el conjunto de genomas analizados (e.g. $D_{a,i}(b)$, $D_{a,i}(c)$, $D_{a,i}(d)$,..., $D_{a,i}(n)$); el conjunto de n POs incluidos en este análisis es solo una muestra tomada del universo completo de POs que se obtendría si se tomaran en cuanta todos genomas procariotes en el planeta. El modelo también toma en cuenta la posibilidad de que el gene de referencia $G_{a,i}$ no tenga un xenólogo entre el conjunto de POs, es decir la hipótesis nula, y puede interpretarse como la probabilidad de que todos los POs son en realidad ortólogos, que no se ha secuenciado todavía el genoma que posee el verdadero xenólogo, o bien que existió una transferencia ancestral hace mucho tiempo y los genes han tenido la oportunidad de divergir significativamente.

Si $G_{a,i}$ es el gene referencia, entonces la hipótesis de que el verdadero xenólogo de $G_{a,i}$ está en el genoma b, es decir que $G_{a,i}$ y $G_{b,j}$ son GCXs, se puede expresar mediante la relación:

$$H(G_{a,i}) = \begin{cases} b & \text{Si } G_{b,j} \text{ es el xenólogo de } G_{a,i} \\ 0 & \text{Si } G_{a,i} \text{ no tiene xenólogos en la muestra de POs} \end{cases}. \quad (1.7)$$

La hipótesis nula puede escribirse como $H(G_{a,i})=0$. Se asume que las diferencias de CRI entre cualquier par de xenólogos recientes siguen una distribución normal con media cero y una desviación estándar muy pequeña σ_h. Esto es, para toda hipótesis $H(G_{a,i})=b$ las diferencias $D_{a,i}(b)$ se distribuyen de la siguiente manera:

$$D_{a,i}(b) \mid H(G_{a,i}) = b \sim \frac{1}{\sigma_h \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{D_{a,i}(b)}{\sigma_h}\right)^2}. \quad (1.8)$$

De manera similar, si los POs $G_{a,i}$ y $G_{b,j}$ no son xenólogos, $H(G_{a,i})=0$, se asume que las diferencias $D_{a,i}(b)$ siguen una distribución normal con media cero pero una desviación estándar σ_o mucho más grande que σ_h ($\sigma_o >> \sigma_h$):

$$D_{a,i}(b) \mid H(G_{a,i}) = 0 \sim \frac{1}{\sigma_o \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{D_{a,i}(c)}{\sigma_o}\right)^2} \quad (1.9)$$

Para facilitar el desarrollo matemático, la constante $\frac{1}{\sqrt{2\pi}}$ no se tomará en cuenta de aquí en adelante —al derivar la ecuación del modelo, se incorporará este valor a la constante global de normalización.

En términos Bayesianos se desea calcular la probabilidad posterior de la hipótesis que propone que $G_{a,i}$ y $G_{b,j}$ son GCXs, dado el vector $\overline{D}$ de las diferencias de CRI, con respecto al genoma a, entre $G_{a,i}$ y sus n POs, esto es, $P(H(G_{a,i}) = b \mid \overline{D})$ donde $\overline{D} = \left[D_{a,i}(b), D_{a,i}(c) \ldots D_{a,i}(n)\right]$. De acuerdo a la Ecuación 1.4, esta probabilidad se puede escribir como

$$P(H(G_{a,i}) = b \mid \overline{D}) = \frac{P(\overline{D} \mid H(G_{a,i}) = b) * P(H(G_{a,i}) = b)}{\sum_{c=0}^{n} P(\overline{D} \mid H(G_{a,i}) = c) * P(H(G_{a,i}) = c)}. \quad (1.10)$$

Donde $P(\overline{D} \mid H(G_{a,i}) = b)$ es el modelo que describe el comportamiento estadístico de las diferencias de CRI, dada la hipótesis de una transferencia horizontal del gene $G_{a,i}$ entre los genomas a y b. $P(H(G_{a,i})=b)$ representa el conocimiento *a priori* que se tiene en cuanto a la posibilidad de que $G_{b,j}$ es el xenólogo verdadero de $G_{a,i}$. Para realizar los cálculos, se asumió que

las diferencias en CRI entre POs son independientes. Evaluaciones teóricas y empíricas al supuesto de independencia han concluido que tiene un desempeño extremadamente eficiente [126]. Si el gene $G_{a,i}$ tiene n POs y proponemos la hipótesis $H(G_{a,i})=b$, el modelo que describe el comportamiento de las diferencias de CRIs se puede obtener aplicando la Ecuación 1.6:

$$P(\overline{D} \mid H(G_{a,i}) = b) \alpha \frac{1}{\sigma_h} e^{-\frac{D_{a,i}(b)^2}{2\sigma_h^2}} \prod_{\substack{c=1 \\ c \neq b}}^{n} \frac{1}{\sigma_o} e^{-\frac{D_{a,i}(c)^2}{2\sigma_o^2}} . \tag{1.11}$$

Como $\overline{D}$, σ_h y σ_o son conocidos al momento evaluar la hipótesis $H(G_{a,i})=b$, se puede introducir una constante adicional para simplificar la Ecuación 1.11:

$$C_1 = \prod_{c=1}^{n} \frac{1}{\sigma_o} e^{-\frac{D_{a,i}(c)^2}{2\sigma_o^2}} . \tag{1.12}$$

La constante C_1 incluye el factor adicional $\frac{1}{\sigma_o} e^{-\frac{D_{a,i}(b)^2}{2\sigma_o^2}}$ que está ausente en la Ecuación 1.11 —nótese que si se toma en cuenta el factor $\frac{1}{\sqrt{2\pi}}$ de las Ecuaciones 1.7 y 1.8, sería obvio incorporarlo a C_1 como $\left(\frac{1}{\sqrt{2\pi}}\right)^n$. Por lo tanto, quitando el signo de proporcionalidad, la Ecuación 1.11 puede rescribirse como:

$$P(\overline{D} \mid H(G_{a,i}) = b) = \frac{\frac{1}{\sigma_h} e^{-\frac{D_{a,i}(b)^2}{2\sigma_h^2}}}{\frac{1}{\sigma_o} e^{-\frac{D_{a,i}(b)^2}{2\sigma_o^2}}} C_1. \tag{1.13}$$

Simplificando algebraicamente la Ecuación 1.13 tenemos entonces,

$$P(\overline{D} \mid H(G_{a,i}) = b) = \frac{\sigma_o}{\sigma_h} e^{-\frac{D_{a,i}(b)^2}{2}\left(\frac{1}{\sigma_h^2} - \frac{1}{\sigma_o^2}\right)} C_1. \tag{1.14}$$

El modelo predictivo final lo podemos derivar al sustituir la Ecuación 1.14 en la Ecuación 1.10:

$$P(H(G_{a,i})=b\mid\overline{D})=\frac{C_1\frac{\sigma_o}{\sigma_h}e^{-\frac{D_{a,i}(b)^2}{2}\left[\frac{1}{\sigma_h^2}-\frac{1}{\sigma_o^2}\right]}P(H(G_{a,i})=b)}{C_1\sum_{c=0}^{n}\frac{\sigma_o}{\sigma_h}e^{-\frac{D_{a,i}(c)^2}{2}\left[\frac{1}{\sigma_h^2}-\frac{1}{\sigma_o^2}\right]}P(H(G_{a,i})=c)}. \quad (1.15)$$

Claramente la constante C_1 se cancela de la ecuación. Para calcular la probabilidad de la hipótesis nula es necesario considerar $\sigma_h=\sigma_o$, porque σ_o representa la distribución de genes no xenólogos, de modo que al sustituir en la Ecuación 1.15 se obtiene:

$$P(H(G_{a,i})=0\mid\overline{D})=\frac{P(H(G_{a,i})=0)}{\omega}. \quad (1.16)$$

Donde $\omega=\sum_{c=0}^{n}\frac{\sigma_o}{\sigma_h}e^{-\frac{D_{a,i}(c)^2}{2}\left[\frac{1}{\sigma_h^2}-\frac{1}{\sigma_o^2}\right]}P(H(G_{a,i})=c)$. La sumatoria empieza en cero para incluir a la hipótesis nula. Desafortunadamente, actualmente no contamos con conocimiento *a priori* completo que nos de pistas de cual es la probabilidad de que los organismos analizados transfieran genes entre ellos, de tal manera que $\sum_{c=1}^{n}P(H(G_{a,i})=c)=1$. Por este motivo, los cálculos se realizaron imparcialmente asumiendo que todos los organismos tienen la misma probabilidad de intercambiar genes. Si se tienen *n* POs y tomando en cuenta a la hipótesis nula, la probabilidad *a priori* de todos los POs a ser GCXs es:

$$P(H(G_{a,i})=b)=\frac{1}{n+1} \quad (1.17)$$

El cálculo de la constante de normalización ω lo haremos por partes. Primero, aplicando la ecuación 1.5 a nuestros datos y tomando en cuenta la hipótesis nula, tenemos que

$$\omega=\sum_{c=0}^{n}P(\overline{D}\mid H(G_{a,i})=c)P(H(G_{a,i})=c). \quad (1.18)$$

El siguiente paso es sustituir términos partiendo de las ecuaciones 1.14 y 1.16 (no se incluye la Ecuación 1.17 para conservar el modelo más general):

$$\omega=P(H(G_{a,i})=0)+\sum_{c=1}^{n}\frac{\sigma_o}{\sigma_h}e^{-\frac{D_{a,i}(c)^2}{2}\left[\frac{1}{\sigma_h^2}-\frac{1}{\sigma_o^2}\right]}P(H(G_{a,i})=c). \quad (1.19)$$

Podemos ahora definir una constante $K = 1/\omega$ de tal manera que el modelo final de predicción quedaría como

$$P(H(G_{a,i}) = b \mid \overline{D}) = K \frac{\sigma_o}{\sigma_h} e^{-\frac{D_{a,i}(b)^2}{2}\left[\frac{1}{\sigma_h^2} - \frac{1}{\sigma_o^2}\right]} P(H(G_{a,i}) = b). \qquad (1.20)$$

De acuerdo con la Ecuación 1.17, todo $PO(G_{a,i},b)$ tiene la misma probabilidad de ser GCX de $G_{a,i}$. Este es un caso particular propiciado por la falta de información pero permite simplificar la Ecuación 1.20. Para efectos prácticos, el modelo empleado se simplificó al sustituir la Ecuación 1.17 en la ecuación 1.19:

$$\omega = \frac{1}{(n+1)}\left[1 + \sum_{c=1}^{n} \frac{\sigma_o}{\sigma_h} e^{-\frac{D_{a,i}(c)^2}{2}\left[\frac{1}{\sigma_h^2} - \frac{1}{\sigma_o^2}\right]}\right]. \qquad (1.21)$$

La probabilidad *a priori* $P(H(G_{a,i})=b)$ de la Ecuación 1.20 también debe sustituirse como lo indica la Ecuación 1.17, y la constante K se modifica ahora tal que $K = \frac{1}{\omega(n+1)}$. El modelo simplificado de predicción final es

$$P(H(G_{a,i}) = b \mid \overline{D}) = K \frac{\sigma_o}{\sigma_h} e^{-\frac{D_{a,i}(b)^2}{2}\left[\frac{1}{\sigma_h^2} - \frac{1}{\sigma_o^2}\right]}. \qquad (1.22)$$

Nótese que el término $(n+1)$ de la constante K se cancela con el término $(n+1)$ de la constante ω, simplificando aun más la Ecuación 1.21. Es importante enfatizar que la Ecuación 1.22 es sólo un caso particular del modelo general presentado en la Ecuación 1.20.

Para efecto de los cálculos, la probabilidad posterior de la hipótesis nula se utilizó como umbral para definir cuando un par de genes es considerado GCXs. La selección de GCXs se llevó a cabo definiendo conservadoramente σ_h=0.002804 y σ_o=0.1, implicando que una diferencia de CRI es considerada significativa si $|D_{a,i}(b)|$<0.0075 (Nótese que esta diferencia representa aproximadamente 1/18 o menos de la escala del CRI, ver Figuras 1.3 y 1.4).

Capítulo II

Uso de codones típico: una zona de tolerancia para alcanzar niveles adecuados de expresión

Nos estamos ahogando en información, pero sufrimos de inanición por falta de conocimiento.

JOHN NAISBETT

2.1 Resumen del capítulo

En el Capítulo I se propuso la existencia de una zona de tolerancia, o seguridad, donde mostrar una composición típica de codones es suficiente para que genes importados sean traducidos eficientemente y puedan alcanzar niveles apropiados de expresión. El Capítulo II presenta evidencias adicionales en favor de esta hipótesis. Tomando como referencia a *Escherichia coli* K12, se calculó la correlación entre las concentraciones de tRNAs —bajo condiciones diferentes de crecimiento—, el UC y el uso de aminoácidos. Los resultados obtenidos sugieren fuertemente que, en efecto, mostrar un UC típico es suficiente para obtener una correlación significativa con las concentraciones de tRNA. Además, contrario a lo que podría esperarse, los genes altamente expresados (GAEs) tienen principalmente una composición típica de codones y no exhiben, en comparación con otros genes de menor expresión, la correlación más elevada con las abundancias de tRNA. La relevancia de estos resultados motivó la evaluación, desde una perspectiva independiente a la transferencia horizontal, de los supuestos básicos que

relacionan al uso de codones (UC), con la *traducibilidad*[10] (eficiencia de la traducción) y el *nivel de expresión*[11].

Por medio de un estudio comparativo de 148 genomas procariotes secuenciados a la fecha, se observó que si bien algunos GAEs muestran un UC rico en un genoma referencia, es incorrecto asumir que sus ortólogos en otros organismos también tienen un UC rico y que su correlación con las concentraciones de tRNA será tan buena como en el genoma referencia. Todo indica que se han sobreinterpretado los resultados de los estudios iniciales del UC realizados en organismos modelo. Por lo tanto, no es posible predecir el nivel de expresión de los genes únicamente a partir del UC, y los métodos que así lo han hecho hasta ahora manejan argumentos circulares llegando a resultados poco confiables. La naturaleza del UC es tal que los genes con un UC "óptimo" son diferentes para cada especie, presumiblemente debido a una combinación de factores estocásticos y restricciones nutricionales relacionadas con el nicho ecológico de cada organismo.

2.2 Antecedentes

Hace ya algunas décadas, se observó que en genes de eucariotes y procariotes la selección de codones sinónimos no es azarosa [127-129]. La mayoría de los genes en un organismo, si no es que todos, muestran un sesgo, algunas veces sutil pero siempre estadísticamente significativo, hacia seleccionar las mismas opciones de codones sinónimos para codificar algún aminoácido particular. El cúmulo cada vez mayor de evidencias a favor de esta observación, permitió sugerir que el genoma debe ser en realidad la unidad operativa objeto de selección y no los genes individuales. Esta idea dio paso al planteamiento de la "hipótesis del genoma" [22], la cual postula que cada gene tiende a seguir el patrón de UC del genoma al que pertenece. Con seguridad no existe un único proceso selectivo que sea responsable por las preferencias de codones observadas en los genomas, dado que una variedad de restricciones biológicas, más allá de codificar para un péptido, pueden actuar sobre el mRNA (e.g. procesamiento, estructura secundaria, estabilidad, propiedades del DNA reflejadas en el mRNA como la susceptibilidad al daño mutagénico y señales relacionadas con la replicación, etc.). Inclusive, antes de que se

[10] Número de proteínas traducidas a partir de un mensajero.
[11] Producción neta de proteína a partir de un gene.

obtuviera la secuencia del primer gene, ya se habían sugerido varias razones del porque la selección natural podría favorecer un codón sinónimo sobre otro [130].

A principios de los 80s se estableció formalmente la existencia de una correlación positiva entre el UC de los genes altamente expresados (GAE) y la concentración de tRNAs, primero en el organismo modelo *Escherichia coli* [23] y posteriormente en levadura [24, 131]. Algunas interpretaciones de este fenómeno extendieron el papel biológico de los codones raros, que en la célula codifican para especies escasas de tRNA, llegando a proponer funciones relacionadas con regulación; por ejemplo, en la modulación de la expresividad de los genes dentro de un mismo operón [132], o directamente como una estrategia evolutiva para mantener los niveles bajos de expresión de algunos genes [131, 133, 134]. Sin embargo, esta última hipótesis fue rebatida por evidencia indicando que la enzima cloramfenicol acetiltransferasa (CAT), que tiene un 25% de codones raros, puede alcanzar altos niveles de expresión [135], y si todavía se le insertan más codones raros, es posible detectar efectos en la traducción únicamente a muy elevadas tasas de transcripción [25]. Estudios posteriores encontraron que no existe una clara preferencia de codones raros en genes reguladores, sugiriendo así que el uso de codones poco abundantes refleja principalmente una relajación de la selección natural [136]. Por otro lado, se encontró evidencia experimental sugiriendo que la velocidad de traducción entre genes es variable y se propuso que la elongación durante la traducción es el factor limitante en la tasa de traducción, argumentando que el grado en que el ribosoma hace más lento su movimiento es proporcional al inverso de la concentración de tRNA [137]. En apoyo a estos resultados se observó que en atenuadores el uso de codones raros contiguos puede generar pausas en la traducción de péptidos líderes [138, 139], aunque aparentemente estas pausas, o variaciones en la velocidad de traducción de ciertos codones, no están linealmente correlacionadas con la concentración de tRNAs [140]. En contraposición con estos argumentos se ha observado, en una cepa silvestre de *E. coli*, que el represor lac (*lacI*), rico en codones raros, se puede llegar a traducir aproximadamente a la misma velocidad de las proteínas ribosomales [141]. De hecho, se ha reportado que las tasas de elongación de polipéptidos son relativamente constantes y que un UC óptimo no es necesario para obtener tasas de elongación razonables [142]. De ser así, esto implicaría que las diferencias en expresión de los genes de *E. coli* no son una consecuencia directa de su variabilidad en UC, es decir, aunque el UC puede estar modulado para obtener una mejor traducibilidad, y a su vez está

relacionado con la expresión, no es posible afirmar que el nivel de expresión de los genes es modulado por el UC o viceversa.

El conjunto de evidencias que minan la fuerza predictiva de la relación entre el uso de codones y el nivel de expresión [140-142] paso en gran parte desapercibido. Resultaba mucho más atractiva la idea de conocer una propiedad fisiológica partiendo únicamente de la secuencia de DNA. Por lo tanto, se siguió argumentando que la correlación entre el UC y el nivel de expresión es suficientemente elevada como para que una variable puede predecir a la otra, tanto en levadura [29, 143] como en *E. coli* [30]. Como consecuencia, la hipótesis que busca explicar la correlación entre UC y niveles de expresión postula que el elevado *sesgo en UC*[12] de los GAEs refleja más claramente el rastro que ha dejado la selección por una mayor traducibilidad, al favorecer el aprovechamiento óptimo de las concentraciones de tRNA para alcanzar niveles máximos de expresión, de tal suerte que todo gene con UC similar al de los GAEs debe ser también altamente expresado. Sin embargo, esta hipótesis se basa en dos supuestos muy debatibles: (1) que el uso de aminoácidos (UAA) está principalmente determinado por restricciones funcionales y estructurales, por lo tanto, es poco o nada informativo en cuanto a la traducibilidad; y (2) que aquellos genes con UC similar al de los GAEs son también transcritos en altas concentraciones (o producen mensajeros muy estables). Estos supuestos no consideran que la redundancia del código genético estándar, sumada a la del segundo código genético (el que establece la correspondencia entre aminoácidos y la estructura tridimensional de la proteína) proporcionan suficientes grados de libertad como para mejorar u optimizar, de ser necesario, la traducibilidad de los genes. De hecho, han surgido más evidencias ilustrando que existen proteínas altamente expresadas con un sesgo relativamente bajo de UC [144], pero se argumentó que estos casos pueden ser las excepciones que confirman la regla. También se ha reportado que un UC óptimo es seleccionado para maximizar la *exactitud de la traducción*[13] [145] y que la selección por una mayor traducibilidad también tiene un impacto significativo en la composición de aminoácidos [93, 94]. En suma, es probable que tanto la eficiencia como la exactitud de la traducción sean factores importantes que ejercen una mayor influencia sobre el UC y el UAA que el nivel de expresión.

[12] Tendencia a usar preferentemente un sólo codón sinónimo por aminoácido.
[13] Incorporación del aminoácido correcto en la cadena polipeptídica naciente por cada codón en el mRNA.

Entre el conjunto de genes particularmente atractivos por mostrar altos niveles de expresión destacan notablemente las proteínas ribosomales (PRs), proponiéndose que los genes con UC parecido al de las PRs (y a algunos otros genes muy abundantes como proteínas de membrana y factores de traducción), deben ser también altamente expresados. Partiendo de este principio, diversas metodologías buscaron medir y/o predecir el nivel de expresión de los genes, basándose exclusivamente en su UC [29, 30, 86, 104, 131, 132, 146, 147]. Este principio de predicción se sigue aplicando actualmente [90-92, 148, 149], a pesar de que se han encontrado especies donde los genes de alta y baja expresión no parecen diferir en su UC [150-152]. Si el UC de los GAEs está en verdad optimizado, debe demostrarse claramente (1) que entre GAEs existe poca variabilidad de UC; (2) que los GAEs exhiben, en comparación con otros genes de menor expresión, la mas alta correlación con la concentración de tRNA; y (3) que los GAEs muestran preferentemente un UC rico en todos los genomas.

Debido a que nos encontramos sumergidos en la era genómica, ahora es posible, gracias a una mayor cantidad de información disponible y a la existencia de evidencias en principio contradictorias, evaluar los conceptos que constituyen las bases metodológicas de los análisis bioinformáticos modernos, con el objetivo de verificar su validez y robustez ante el incremento exponencial de datos producto de los proyectos genómicos. Este capítulo demuestra que un UC típico es suficiente para alcanzar una correlación significativa con la concentración de tRNA y, complementando el estudio con el análisis comparativo de 148 genomas disponibles públicamente, se cuestiona la interpretación inicial de las observaciones que dieron origen a los métodos clásicos de predicción de niveles de expresión, proponiéndose una interpretación más congruente con los datos actuales.

2.3 Objetivo

Determinar si el UC típico define una zona de tolerancia o seguridad donde los genes pueden alcanzar niveles adecuados de expresión. De ser así, necesariamente el uso de codones genómico debe correlacionar significativamente con las abundancias de tRNA. En principio, un resultado como sería incompatible con el paradigma vigente que permite predecir niveles de expresión a partir únicamente del UC —todo gene con preferencias de codones similares a genes cuyo producto proteínico es muy abundante, son también altamente expresados. Por lo tanto, es necesario hacer una evaluación de los supuestos esenciales que subyacen detrás de este

paradigma. Específicamente se buscará dar respuesta a las siguientes preguntas: (1) ¿Es cierto que el UC de los GAEs muestra la más alta correlación con las abundancias de tRNAs? O en otras palabras ¿Existen genes que no sean considerados como altamente expresados y que muestren una mejor correlación con la concentración de tRNA? (2) Comúnmente se asume que la composición de codones en GAE es óptima para la traducción en genomas que muestran altas tasas de crecimiento, ¿en verdad es este un fenómeno universal, que permita la predicción, con un nivel de confianza aceptable, de GAE en otros organismos donde se conoce poco o nada sobre niveles de expresión y concentraciones de tRNA?

Contestar la primera pregunta requiere determinar, para un organismo modelo como *E. coli*, si al menos durante la fase de crecimiento exponencial el contenido de tRNA correlaciona efectivamente mejor con el UC de los GAEs en comparación al resto del genoma. *E. coli* es la bacteria modelo que vio nacer los supuestos que en este trabajo se someten a prueba y, por lo tanto, es un organismo excelente para hacer el análisis. Además, ya se han reportado las concentraciones de tRNA para *E. coli* a diferentes tasas de crecimiento [153]. Por otro lado, la segunda pregunta requiere de un enfoque comparativo para determinar si existe un conjunto de genes altamente expresados (e.g. proteínas ribosomales) común a todos los organismos procariotes reportados a la fecha y que muestre un UC preferentemente rico. La existencia de un conjunto de genes con estas características implicaría que efectivamente se puede hablar de una optimización universal del UC en algunos genes, los cuales constituirían la mejor referencia para medir traducibilidad —no necesariamente niveles de expresión.

Las Secciones 2.5, y 2.6 dan respuesta a la primera pregunta. La Sección 2.7 presenta una evidencia adicional en apoyo a la relación UAA–traducibilidad y por lo tanto justifica la influencia de la composición de aminoácidos en el cálculo del CRI. La Secciones 2.8 y 2.9 dan respuesta a la segunda pregunta.

2.4 Hipótesis

El UC de codones genómico correlaciona de manera significativa con la concentración de tRNAs, permitiendo que la mayoría de los genes en un organismo sean traducidos eficientemente. En otras palabras, la célula no está especializada en traducir de manera óptima un número pequeño de genes altamente expresados (e.g. aquellos relacionados con el proceso de traducción) que le permiten alcanzar máximas tasas de crecimiento. En vida libre, la bacteria se encuentra la

mayor parte de su vida en condiciones donde los nutrimentos son escasos (fase estacionaria), por lo tanto, genes que le permitan sobrevivir bajo una variedad de estreses deben ser traducidos eficientemente y en niveles apropiados, cuando se les necesita, para garantizar su supervivencia. En términos evolutivos, la selección natural ha guiado la célula hacia un equilibrio tal que el UC en la mayoría de los genes correlaciona bien con el contenido de tRNAs, permitiéndoles alcanzar tasas de traducción y niveles de expresión adecuados. Al menos dos implicaciones de esta hipótesis serían: (1) un gene con un UC tal que pueda explotar óptimamente la disponibilidad de tRNAs no significa que es altamente expresado, únicamente que es traducido con alta eficiencia, independientemente de la cantidad de producto proteico generada; y (2) que el UC genómico es una buena referencia para medir traducibilidad.

2.5 El UC genómico correlaciona mejor con la concentración de tRNA que el UC en proteínas ribosomales

En esta sección y la siguiente se buscará responder a la primera pregunta planteada en los objetivos (Sección 2.3). De ser cierto que el UC de los GAE está optimizado para explotar de la manera más eficiente las abundancias de tRNAs en la célula, propiedad que se presume les permite ser altamente expresados, debería ser poco probable encontrar un conjunto de genes de baja expresión que mostraran una mejor correlación con la concentración de tRNA. Gracias a la acumulación de datos experimentales, ahora es factible someter a prueba esta deducción.

En 1996 Dong y colaboradores [153] midieron, en *E. coli* K12, la concentración de tRNA para 5 velocidades de crecimiento (Tabla 2.1). Los autores también reportaron las frecuencias de codones en el conjunto mRNAs transcritos para un total de 190 genes altamente expresados (ver Tabla 2.2), que comprenden dos tercios de la masa proteínica de la célula —la abundancia de cada gene es mayor al 0.05% del total de proteínas en al menos una condición [154, 155]. La Tabla 2.3 muestra el anticodón y los codones que traduce cada tRNA reportados en la literatura [143, 156-159]. Con esta información es posible calcular la correlación entre el UC de distintos grupos de genes y las abundancias de tRNAs. Si dos o más tRNAs traducen un mismo codon, la frecuencia total de dicho codon es repartida entre los tRNAs involucrados en base a sus concentraciones relativas, pues se asume que el tRNA más abundante tendrá más posibilidades de traducir el codón. Si un tRNA traduce 2 codones, la suma de las frecuencias de ambos codones se correlaciona con la concentración del tRNA respectivo.

Tabla 2.1. La concentración intracelular (μM) de tRNA medida a diferentes tasas de crecimiento.

tRNA	Velocidad de crecimiento (duplicaciones / hora)				
	0.4	**0.7**	**1.07**	**1.6**	**2.5**
Ala1B	10.25	11.73	14.06	17.52	20.97
Ala2	1.95	2.12	2.33	3.19	3.57
Arg2	15.00	14.54	15.54	23.77	25.57
Arg3	2.01	2.61	1.45	2.26	2.30
Arg4	2.74	2.35	2.64	3.26	3.52
Arg5	1.23	1.57	1.61	2.46	2.20
Asn	3.77	3.86	4.35	6.10	7.29
Asp1	7.56	8.13	8.42	12.04	15.46
Cys	5.01	4.88	5.23	7.04	7.07
Gln1	2.41	2.72	3.63	3.17	4.38
Gln2	2.78	3.08	3.47	5.07	6.27
Glu2	14.88	15.58	16.71	24.12	29.35
Gly1_2	6.75	7.18	7.74	10.95	11.08
Gly3	13.76	15.21	16.75	19.84	24.96
His	2.02	2.19	2.63	3.35	4.38
Ile1_2	10.96	11.85	13.24	18.92	24.74
Leu1	14.11	14.91	16.76	21.32	22.20
Leu2	2.97	3.47	4.04	4.72	5.93
Leu3	2.10	2.49	2.62	3.19	3.17
Leu4	6.04	6.33	6.97	9.66	9.30
Leu5	3.57	3.47	4.07	3.65	3.78
Lys	6.08	6.80	7.35	8.73	10.43
Met_f1	3.82	4.82	6.01	6.33	10.22
Met_f2	2.26	2.28	2.36	3.38	3.77
Met_m	2.23	2.59	2.91	4.10	4.43
Phe	3.27	3.60	4.29	4.69	5.11
Pro1	2.84	2.44	3.51	2.75	2.67
Pro2	2.27	2.51	2.26	4.01	3.75
Pro3	1.83	1.89	2.22	2.55	2.56
Ser1	4.09	5.56	5.47	6.98	7.36
Ser2	1.09	1.04	1.17	1.37	1.45
Ser3	4.44	4.39	4.53	5.40	5.67
Ser5	2.41	2.60	2.87	3.68	4.03
Thr1	0.32	0.41	0.54	0.56	0.67
Thr2	1.71	2.00	2.11	2.67	3.12
Thr3	3.46	3.73	3.87	4.86	5.54
Thr4	2.89	3.17	3.25	4.99	6.89
Trp	2.98	2.78	3.35	4.15	5.02
Tyr1	2.43	2.41	2.70	4.61	4.19
Tyr2	3.98	3.86	3.75	5.22	5.04
Val1	12.12	12.07	11.07	18.99	20.39
Val2A	1.99	2.00	2.38	2.70	2.79
Val2B	2.00	2.39	2.64	3.61	4.42

Dong y colaboradores [153] calcularon las concentraciones de tRNA a partir del número de ribosomas por célula a diferentes tasas de crecimiento[160], el volumen de células a diferentes tasas de crecimiento [161] y sus propios datos de las razones tRNA/ribosoma [153]. Estos valores son promedios de 6 mediciones con desviación estándar entre ±5% y ±15% [153].

Tabla 2.2. Frecuencias de codones en la población de mRNAs transcritos durante 5 velocidades de crecimiento en *E. coli* K12.

Codón	tRNA	Frecuencia de codones del mRNA pora cada tasa de crecimiento (duplicaciones / hora)				
		0.4	0.7	1.07	1.6	2.5
GGG	Gly1	4.81	4.26	3.57	2.79	2.36
GGA	Gly2	2.71	2.49	2.21	1.79	1.26
GGU	Gly3	38.29	39.18	40.49	42.27	45.55
GGC	Gly3	35.62	35.58	35.54	35.49	34.17
GAG	Glu2	16.57	16.78	17.04	17.31	16.97
GAA	Glu2	53.10	53.94	55.10	56.68	57.86
GAU	Asp1	24.25	23.43	22.40	21.08	19.27
GAC	Asp1	28.72	29.65	30.93	32.35	33.74
GUG	Val1	21.40	20.34	18.93	17.74	14.98
GUA	Val1	15.87	17.05	18.65	19.95	22.31
GUU	Val1	31.31	33.10	35.63	38.14	43.18
GUC	Val2A	11.25	10.58	9.71	8.86	7.67
GCG	Ala1B	30.33	29.55	28.45	27.29	24.11
GCA	Ala1B	22.13	22.19	22.38	23.07	24.87
GCU	Ala1B	28.85	30.31	32.41	34.79	39.49
GCC	Ala2	19.80	18.50	16.81	14.67	11.81
AGG	Arg5	0.09	0.07	0.05	0.03	0.03
AGA	Arg4	1.12	0.99	0.84	0.65	0.63
AGU	Ser3	3.99	3.55	3.01	2.38	2.19
AGC	Ser3	11.97	11.40	10.69	9.88	9.31
AAG	Lys	12.08	12.76	13.74	14.89	17.22
AAA	Lys	44.43	46.41	49.07	51.99	55.01
AAU	Asn	9.79	8.88	7.79	6.43	5.61
AAC	Asn	27.95	28.22	28.64	29.02	29.21
AUG	Met	22.37	22.36	22.34	22.30	21.67
AUA	Ile2	0.93	0.85	0.75	0.61	0.52
AUU	Ile1	21.38	20.45	19.26	17.72	15.79
AUC	Ile1	36.68	37.72	39.15	41.38	43.86
ACG	Thr2	7.53	6.95	6.21	5.20	4.17
ACA	Thr4	3.48	3.25	2.99	2.63	2.61
ACU	Thr1	13.88	15.10	16.76	18.31	20.64
ACC	Thr1	26.51	26.77	27.10	27.47	26.70
UGG	Trp	9.76	9.28	8.69	8.01	7.03
UGA	Stop	0.31	0.27	0.23	0.17	0.19
UGU	Cys	4.23	3.97	3.64	3.24	2.76
UGC	Cys	5.29	5.06	4.77	4.35	3.81
UAG	Stop	0.00	0.00	0.00	0.00	0.00
UAA	Stop	2.77	3.02	3.38	3.65	4.18
UAU	Tyr1	10.68	9.90	8.90	7.85	6.72
UAC	Tyr1	16.20	16.41	16.71	16.90	16.52
UUG	Leu4	6.63	6.22	5.72	4.93	4.27
UUA	Leu5	6.13	5.46	4.64	3.56	2.73

Codón	tRNA	Frecuencia de codones del mRNA pora cada tasa de crecimiento (duplicaciones / hora)				
		0.4	0.7	1.07	1.6	2.5
UUU	Phe	12.55	11.54	10.30	8.72	7.92
UUC	Phe	22.68	22.55	22.44	22.68	23.25
UCG	Ser1	6.05	5.41	4.58	3.75	2.51
UCA	Ser1	3.89	3.54	3.09	2.55	1.98
UCU	Ser1	13.12	13.54	14.14	14.84	16.33
UCC	Ser5	11.15	11.57	12.09	12.34	11.68
CGG	Arg3	1.75	1.52	1.23	0.90	0.62
CGA	Arg2	1.32	1.17	0.99	0.75	0.67
CGU	Arg2	31.12	33.46	36.61	39.60	43.82
CGC	Arg2	22.25	22.31	22.39	21.76	20.59
CAG	Gln2	29.24	28.80	28.33	27.69	27.28
CAA	Gln1	10.19	9.65	8.98	7.99	7.01
CAU	His	9.23	8.72	8.11	7.23	6.78
CAC	His	13.90	13.90	13.91	14.08	14.21
CUG	Leu1	60.13	60.62	61.29	61.59	60.75
CUA	Leu3	2.15	1.87	1.53	1.09	0.82
CUU	Leu2	5.70	5.22	4.64	4.01	3.86
CUC	Leu2	6.19	5.91	5.52	5.03	4.09
CCG	Pro1	29.51	29.22	28.88	28.91	28.82
CCA	Pro3	6.52	6.47	6.40	6.08	5.18
CCU	Pro2	4.99	4.90	4.79	4.62	4.38
CCC	Pro2	3.32	2.77	2.10	1.40	1.09

Tabla 2.3. tRNAs en *Escherichia coli*, su anticodón y codones que reconoce.

tRNA	Anticodón (5'→3'')	Codones reconocidos (5'→3')
Ala1B	UGC	GCU GCA GCG
Ala2	GGC	GCC
Arg2	ACG	CGU CGC CGA
Arg3	CCG	CGG
Arg4	UCU	AGA
Arg5	CCU	AGG
Asn	GUU	AAC AAU
Asp1	GUC	GAC GAU
Cys	GCA	UGC UGU
Gln1	UUG	CAA
Gln2	CUG	CAG
Glu2	UUC	GAA GAG
Gly1[a]	CCC	GGG
Gly2	UCC	GGA GGG
Gly3	GCC	GGC GGU
His	GUG	CAC CAU
Ile1	GAU	AUC AUU
Ile2[a]	UAU	AUA
Leu1	CAG	CUG

tRNA	Anticodón (5'→3'')	Codones reconocidos (5'→3')
Leu2	GAG	CUC CUU
Leu3	UAG	CUA CUG
Leu4	CAA	UUG
Leu5	UAA	UUA UUG
Lys	UUU	AAA AAG
Met f1	CAU	AUG
Met f2	CAU	AUG
Met m	CAU	AUG
Phe	GAA	UUC UUU
Pro1	CGG	CCG
Pro2	GGG	CCC CCU
Pro3	UGG	CCA CCU CCG
Ser1	UGA	UCA UCU UCG
Ser2	CGA	UCG
Ser3	GCU	AGC AGU
Ser5	GGA	UCC UCU
Thr1	GGU	ACC ACU
Thr2	CGU	ACG
Thr3	GGU	ACC ACU
Thr4	UGU	ACA ACU ACG
Trp	CCA	UGG
Tyr1	GUA	UAC UAU
Tyr2	GUA	UAC UAU
Val1	UAG	GUA GUG GUU
Val2A	GAC	GUC GUU
Val2B	GAC	GUC GUU

Fragmento de la Tabla 2 en Dong et al [153].
[a] Los tRNAs Gly1 y Gly2 son tratados colectivamente al igual que Ile1 e Ile2 por los autores [153].

Tabla 2.4 Para 5 velocidades de crecimiento (columna 1) en *E. coli* K12, se muestran las correlaciones entre la concentración de tRNA y el uso de codones (UC) en 3 conjuntos de genes: mRNAs transcritos (columna 2), todo el genoma (columna 3) y las proteínas ribosomales (PRs) (columna 4).

Tasa de crecimiento (duplicaciones/hora)	Correlaciones‡ entre el contenido de tRNAs y:		
	UC en el mRNA	UC genómico	UC en PRs
0.4	0.8378	0.8200	0.8124
0.7	0.8528	0.8330	0.8209
1.07	0.8581	0.8411	0.8221
1.6	0.8463	0.8310	0.8157
2.5	0.8655	0.8583	0.8311

‡ Todas las correlaciones son significativas ($p < 0.01$) y se calcularon aplicando el factor de correlación de Pearson. Las diferencias en correlación también son significativas (t-test pareado) para cualquier par de columnas en la tabla: mRNA vs Genoma ($p < 0.08$), Genoma vs PRs ($p < 0.05$) y mRNA vs PRs ($p < 0.001$)

Como se ilustra en la Tabla 2.4, las frecuencias genómicas de codones correlacionan mejor con las concentraciones de tRNA que con el UC de las proteínas ribosomales (PRs). Aunque aparentemente la diferencia no es muy grande, sí es estadísticamente significativa ($p < 0.05$). Además, el UC genómico exhibe una correlación consistentemente mayor a través de todas las condiciones de crecimiento analizadas.

Al graficar la concentración de tRNAs contra las frecuencias de codones en el genoma completo y en las PRs (Figura 2.1), se puede apreciar fácilmente que el UC de las PRs muestra una mayor dispersión de la línea de regresión. Para comprender mejor este resultado, la Figura 2.2, panel A, muestra como el $tRNA^{Leu4}$, que traduce el codón raro UUG para el aminoácido abundante leucina, está en mayor concentración que el $tRNA^{His}$ que traduce el codón "óptimo" CAC para el aminoácido escaso histidina. Se puede apreciar también que UUG es más abundante que los dos codones para histidina. Desde el punto de vista de la correlación con la concentración de tRNA, es más ventajoso usar el codón UUG que CAC porque hay más tRNAs que lo traducen. Sin embargo, como se puede apreciar en la Figura 2.2, panel B, si las frecuencias de codones sinónimos se normalizaran por aminoácido, como se practica comúnmente en otras metodologías [30, 92, 162], esta información se perdería totalmente porque entonces el codón óptimo para histidina (CAC) tendría el mismo peso que el codón óptimo para leucina (CUG) y el codon menos frecuente par histidina (CAU) tendría más peso que el codon UUG para leucina. Estos datos (Figuras 1.1 y 1.2) sugieren que las frecuencias genómicas de codones representan una mejor referencia para medir la traducibilidad de los genes que el UC de las PRs, y por lo tanto apoya la propuesta de que un UC típico (zona de seguridad o tolerancia, ver Capítulo I) es suficiente para obtener el nivel de expresión necesarios para que los genes realicen su función adecuadamente.

No es sorprendente que las correlaciones en la Tabla 2.4 no sean tan altas como se había reportado anteriormente (alrededor de 0.95 [137, 142]) cuando se utilizaron mucho menos datos en los análisis. La correlación entre el UC del mRNA producido a cada velocidad de crecimiento y la concentración de tRNA es ligeramente superior (pero significativa; $p < 0.08$) a la del UC genómico (Tabla 2.4). Desafortunadamente, la escasa disponibilidad de datos experimentales sobre concentraciones de mRNA en la mayoría de los genomas obstaculiza la generación de modelos genéricos de análisis que tomen en cuenta esta información.

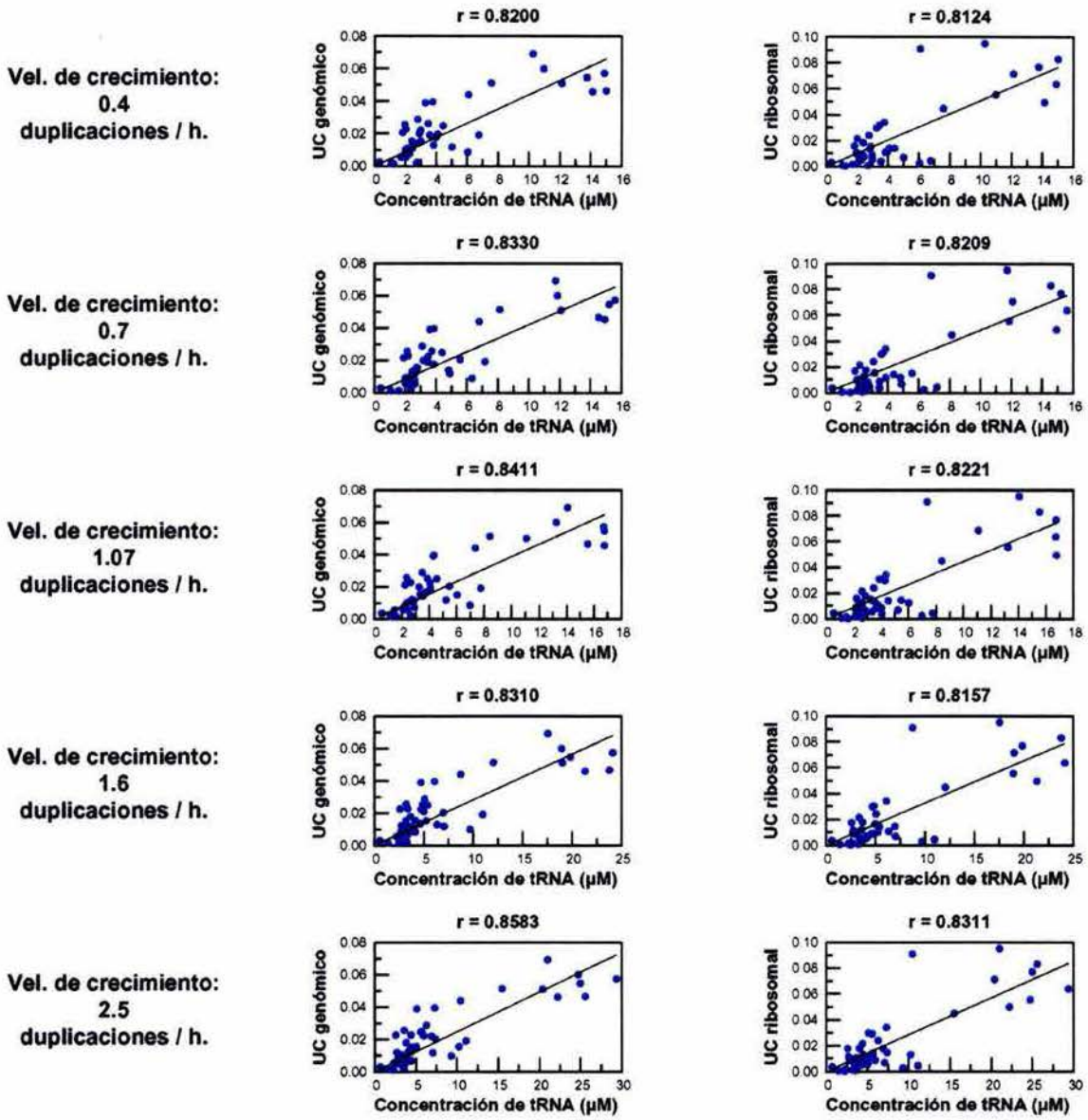

Figura 2.1. Covariación entre las concentraciones de tRNA en *E. coli K12* (eje de las Xs) y el uso de codones (UC) (eje de las Ys) a través de 5 condiciones de crecimiento. Cada columna de paneles corresponde a un grupo de genes: todo el genoma (columna 1) y las proteínas ribosomales (columna 2). Los renglones de paneles corresponden a correlaciones estimadas por cada velocidad de crecimiento. El cálculo de las frecuencias de codones traducidos por cada tRNA se explica en el texto (Sección. 2.5)

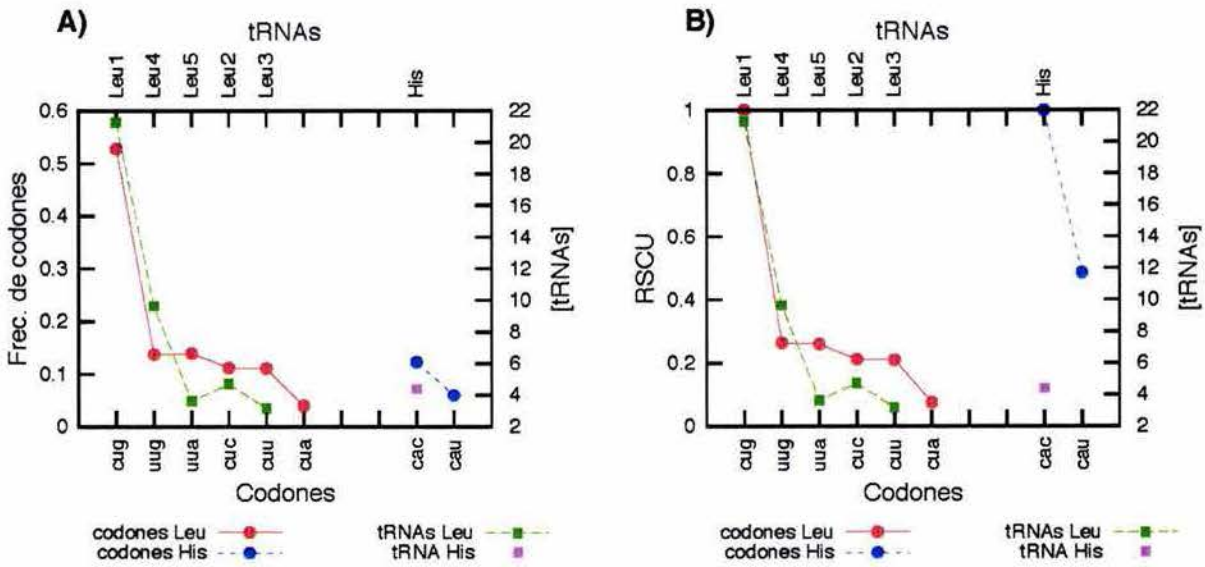

Figura 2.2. Relación entre las frecuencias globales de codones, uso relativo de codones sinónimos, y la concentración de tRNA en *E. coli* K12 creciendo a una velocidad de 1.6 duplicaciones por hora. En el Panel A, el eje X_1 (inferior) muestra los codones sinónimos para leucina (6) e histidina (2). El eje X_2 (superior) muestra los tipos de tRNA que traducen leucina (5) e histidina (1). El eje Y_1 (izquierda) muestra las frecuencias genómicas de codones para leucina (línea roja) y las frecuencias de codones en genes ribosomales que codifican para histidina (línea azul). Las frecuencias de codones se multiplicaron por 10. El eje Y_2 (derecha) muestra las concentraciones de los distintos tRNAs para leucina (línea verde) e histidina (cuadro magenta). Cada tRNA se colocó en la misma posición del codón que traduce preferentemente. En el panel B los ejes X_1, X_2 y Y_2 tienen la misma interpretación que en el panel A, sin embargo, el eje Y_1 ahora presenta las frecuencias relativas de codones sinónimos para leucina (línea roja) e histidina (línea azul). RSCU significa uso relativo de codones sinónimos (por sus siglas en Inglés *Relative Synonymous Codon Usage*) descrito en la Sección 2.6.2. Ver discusión en el texto para la interpretación de la gráfica.

Las correlaciones aquí descritas sugieren que es factible emplear las frecuencias genómicas de codones como referencia para evaluar el grado de armonía entre el UC de genes individuales y la maquinaria de traducción de su genoma. El Índice de Riqueza de Codones o CRI por sus siglas en inglés (*Codon Richness Index*), descrito en el Capítulo I (Sección 1.5), toma ventaja de esta observación para cuantificar el grado en que genes individuales utilizan los codones mas abundantes de su genoma. El razonamiento subyacente detrás de esta propuesta plantea que un genoma debe traducir eficientemente la mayoría de sus genes para sobrevivir, y por ende, el grueso de los genes en el genoma debe correlacionar razonablemente bien con las abundancias de tRNAs.

2.6 Los genes con alto CRI correlacionan mejor con las abundancias de tRNAs que los GAEs

Esta sección complementa la respuesta presentada en la Sección 2.5 a la primera pregunta planteada en los objetivos (Sección 2.3): ¿Existen genes que no sean considerados como altamente expresados y que muestren una mejor correlación con la concentración de tRNA? La identificación del proceso(s) biológico(s) y/o evolutivo(s) que pueda(n) explicar la relación UC–tRNA, requiere de numerosos enfoques teóricos y experimentales. Sin lugar a dudas, uno de ellos es la selección de diversos grupos de genes para estudiar la correlación entre el UC y la concentración de tRNA. Como se encuentra detallado en antecedentes (Sección 2.2), desde varios puntos de vista, estudios de esta naturaleza ya se han llevado a cabo. Sin embargo, la diferencia con el presente análisis radica en los supuestos esenciales empleados, producto de la evaluación de nuevos datos.

Se seleccionaron 6 grupos de genes utilizando criterios funcionales y composicionales: (1) el conjunto de genes predicho/conocido como altamente expresados fue directamente compilado de la literatura [148]; (2) todos los genes anotados en el genoma de *E. coli* K12; (3) Las proteínas ribosomales, ver Sección 2.6.1; (4) Los genes con alto CRI, ver Capítulo I, Sección 1.5.2; (5) Los genes con alto índice de adaptación de codones, o CAI por sus siglas en inglés *Codon Adaptation Index*, ver Sección 2.6.2; y (6) el conjunto de mensajeros producido en cada condición de crecimiento en la que se midió la concentración de tRNA (Tabla 2.2). En la Sección 2.6.3 se presenta y discute la correlación entre la concentración de tRNA y el UC en los 6 grupos de genes antes mencionados.

2.6.1 Obtención de las proteínas ribosomales (PRs)

Aunque de momento sólo se tomarán en cuenta las PRs de *E. coli* K12, más adelante (Sección 2.8) se realizarán comparaciones entre todos los genomas analizados. Por lo tanto, aquí se describe el procedimiento para seleccionar las PRs en todos los genomas. La identificación se realizó inicialmente con base en las anotaciones funcionales reportadas en las bases de dataos. En los casos donde no había PRs anotadas se utilizaron búsquedas con Blast (e-value $\leq 10^{-3}$) contra *E. coli* K12 y *Saccharomyces cerevisiae*, aplicando el criterio de mejor calificación bi-direccional o BDBH descrito en el Capítulo I, Sección 1.7. En este caso, se exigió que el alineamiento cubriera al menos el 70% de la proteína ribosomal anzuelo.

La Figura 2.3 muestra el CRI de todos los genes en *E. coli* K12, resaltando los GAEs (puntos verdes) y las PRs (puntos rojos). Es claro que los GAEs muestran un CRI mayor al promedio genómico, pero están concentrados principalmente en la zona de CRI típico–alto (ver Capítulo I, Sección 1.5.2). Aunque menos claro en la figura, la tendencia es la misma para las PRs. En la Sección 2.8 se discutirán las implicaciones de esta distribución.

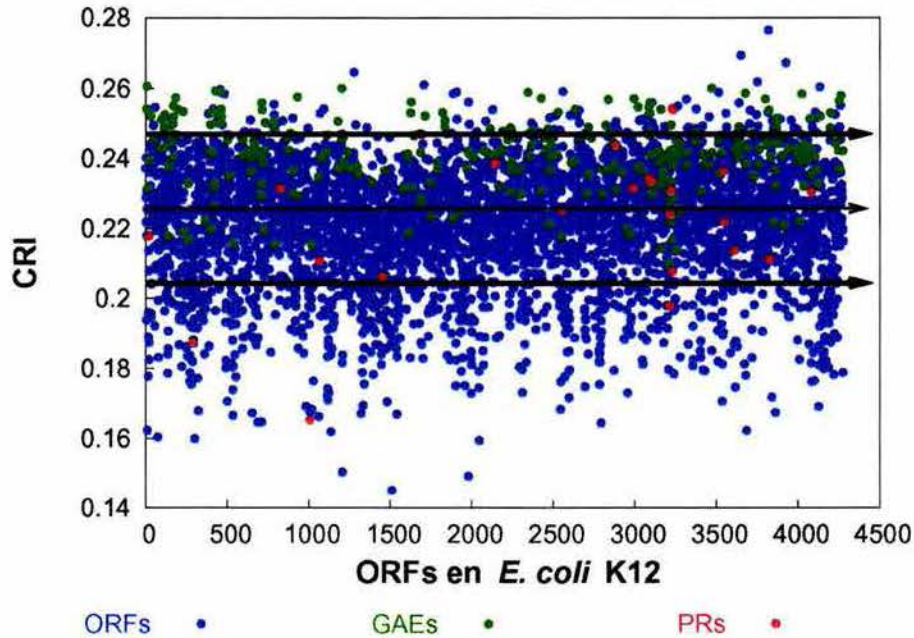

Figura 2.3. El índice de riqueza de codones (CRI) de *E. coli* K12 y sus genes altamente expresados (GAEs). El eje de las Xs presenta los ORFs de acuerdo a su orden de aparición en el cromosoma. La gráfica muestra que las proteínas ribosomales (PRs) y los GAEs [148] tienen en su mayoría un CRI típico-alto y alto. Las flechas indican los diferentes niveles de CRI descritos en el Capítulo I, Sección 1.5.2.

2.6.2 Obtención de genes con alto CAI

El cálculo del CAI [30] requiere que se conozca a priori un conjunto de genes de elevada expresión (principalmente genes relacionados con la traducción), con el propósito de utilizar sus preferencias de codones como referencia para evaluar el UC de genes individuales —todo gene con UC similar al conjunto de entrenamiento (CAI elevado) es entonces predicho como altamente expresado. La Figura 2.4 muestra el CAI de todos los genes en *E. coli* K12 calculado con base en el mismo conjunto de entrenamiento que el reporte original [30]. Como es de esperarse, las PRs muestran valores muy elevados de CAI simplemente porque constituyen el conjunto referencia. Lo mismo sucede con los GAEs pues su definición como GAEs depende de que muestren un CAI

alto (> 0.4). Nótese que no todos los GAEs muestran alto CAI porque fueron predichos con base en una metodología ligeramente diferente pero equivalente [148]. Las 4 PRs con CAI < 0.4 no fueron tomadas en cuenta en la definición original de este índice [30].

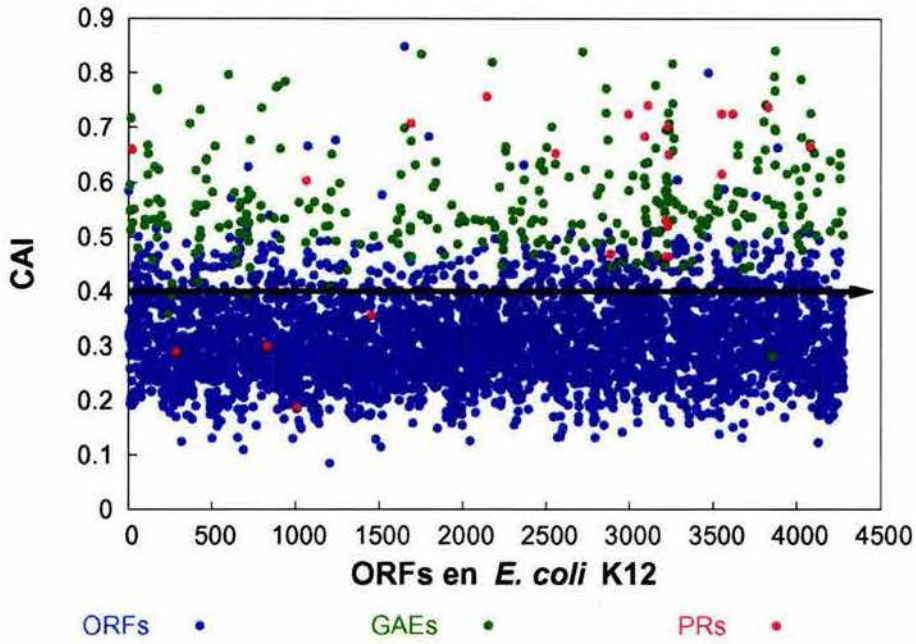

Figura 2.4. El índice de adaptación de codones (CAI) de *E. coli* K12. El eje de las Xs representa los genes de acuerdo a su orden de aparición en el cromosoma. La gráfica muestra que las proteínas ribosomales (PRs) y los genes conocidos/predichos como altamente expresados (GAEs) [148] tienen en su mayoría valores altos de CAI. La flecha horizontal indica el valor de CAI a partir del cual se puede considerar al gene como altamente expresado.

No se estimó el CAI para todos lo genomas porque la metodología presenta tres problemas importantes. Primero, debido a la carencia de evidencias experimentales, no es posible señalar sin ambigüedad cuales son los genes de mayor expresión en la mayoría de los genomas para definir los conjuntos refrencia. Segundo, es necesario asumir no sólo que los ortólogos de genes altamente expresados en organismos modelo son a su vez muy abundantes en el genoma de interés, sino también que el UC en estos genes es óptimo y muestra una elevada correlación con las abundancias —desconocidas— de tRNA. Tercero, el cálculo numérico del CAI presenta otras características no deseables: (a) sólo se toman en cuenta 59 de los 64 codones. Los codones que codifican para Met, Trp y señales de término de la traducción son ignorados bajo el argumento de poca o nula influencia en la eficiencia de la traducción (Sección 2.2). Sin embargo, los codones de paro son señales que participan en la cinética de la traducción al ser interpretadas con distintas

eficiencias por factores de terminación —recordar que la tasa de producción de una proteína es igual a la tasa de terminación de la traducción del mensajero respectivo (ver Capítulo I, Sección 1.5). El CRI considera a los 64 codones como informativos; (b) El CAI se basa en el cálculo de una media geométrica (multiplica la contribución de cada codón y el resultado lo eleva al inverso de la longitud del gene), por lo tanto, para evitar multiplicar por cero cuando un codón no está presente en el conjunto referencia de genes altamente expresados, el CAI le asigna un valor arbitrario de 0.5 como una corrección estadística "razonable" por el tamaño pequeño de la muestra; no obstante, metodologías similares utilizan valores muy diferentes, por ejemplo 0.01 [162]. De no introducir esta corrección, el CAI de todo gene que utilizara este tipo de codones sería cero. El CRI se basa en una media aritmética (la sumatoria de las contribuciones de cada codon dividida por la longitud del gene) y por lo tanto no hay necesidad de introducir correcciones de este tipo, simplemente se asigna una ponderación de 0.0 porque el codón no es utilizado; (c) El CAI aplica el uso relativo de codones sinónimos (RSCU por sus siglas en inglés: *Relative Synonym Codon Usage*), que consiste en normalizar cada codón por la suma de las frecuencias de todos sus codones sinónimos. De esta manera queda descartada por definición toda información relacionada con la composición de aminoácidos. Sin embargo, como se mencionó en antecedentes (Sección 2.2), se ejemplificó en la Sección 2.5 (Figura 2.2) y se muestra más adelante (Sección 2.7), existe una correlación importante entre la concentración celular de tRNAs y el uso de aminoácidos (UAA). El CRI no normaliza los codones sinónimos; y (d) el número de codones AUG y UGG son restados de la longitud total de los genes pues no contribuyen a la calificación final del gene (Met y Trp no tienen codones sinónimos). En este trabajo sí se les considera relevantes para la eficiencia de la traducción, dado que el ribosoma invierte tiempo en traducir estos codones, el cual depende de la concentración de los tRNAs respectivos. En la Tabla 2.1 (Sección 2.5) se puede apreciar como los tRNAs que traducen metionina y triptofano tienen concentraciones mayores y menores que otros tRNAs que traducen aminoácidos con varios codones sinónimos. El CRI y el RLI (Sección 2.7.2) consideran a estos codones tan importantes como sus frecuencias relativas, de manera que no se altera la longitud de los genes. La correlaciones entre los índices CRI, RLI, CAI y AARI se discuten en la Sección 2.7.

2.6.3 Genes con alto CRI muestran la más alta correlación con la concentración de tRNA

La Figura 2.5 muestra, para *E. coli* K12, las correlaciones entre la concentración de tRNAs y el UC en los 6 grupos de genes definidos en la Sección 2.6. Aunque todas las correlaciones son altas y significativas ($r > 0.81$, $p < 0.01$), es evidente que los genes con alto CRI (línea negra continua) están más relacionados con las abundancias de tRNA que el grupo de genes considerados como los más altamente expresados en *E. coli* K12 (líneas punteadas verde y azul), inclusive la correlación es mejor que con el UC del mRNA producido en todas las condiciones de crecimiento estudiadas, confirmando así que el UC genómico es una mejor referencia para medir traducibilidad.

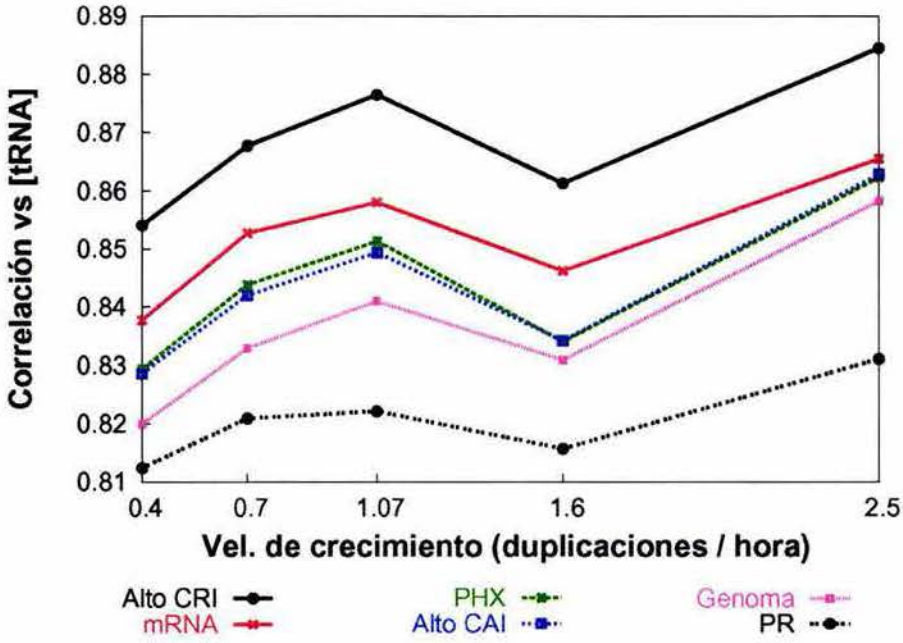

Figura 2.5 Correlación entre la concentración de tRNAs y el uso de codones (UC) en seis grupos de genes: Genes con alto CRI, el mRNA producido por condición de crecimiento, los genes predichos como altamente expresados (PHX) [148], los genes con alto CAI (Sección 2.6.2), todo el genoma y las proteínas ribosomales (PRs). Todas las correlaciones son significativas ($p<0.01$) y se calcularon con el factor de correlación Pearson. Se puede apreciar claramente que el conjunto de genes con alto CRI es el que mejor correlaciona. La gráfica sugiere que el UC genómico es mejor referencia para medir traducibilidad que el UC de las PRs.

En la Figura 2.5 las diferencias observadas entre la correlación de genes con alto CRI y cualquier otro grupo de genes significativas de acuerdo a la prueba estadística *t-test* para muestras pareadas ($p_{max} < 0.05$ y disminuye a $p < 0.01$ para cualquier otro grupo de genes). Sin embargo, las diferencias en correlación entre el mRNA, genes con alto CAI y los genes predichos como altamente expresados (PHX) no son significativas ($p > 0.3$). Esto puede explicarse porque tanto los genes con alto CAI como los PHX están basados en las preferencias de codones de las proteínas ribosomales y conforme aumenta la velocidad de crecimiento aumenta el numero de mensajeros de proteínas ribosomales en la célula. Aunque las correlaciones de los genes con alto CAI y genes PHX son mayores que la de las frecuencias genómicas de codones, las diferencias no son significativas ($p > 0.4$). Como se mencionó en la Sección 2.5 (Tabla 2.4) las diferencias entre el mRNA, el genoma y las PR sí son significativas.

En la Figura 2.3 puede distinguirse la existencia de un número importante (35%) de genes con alto CRI que no son altamente expresados (puntos azules), si estos genes correlacionan mejor con la concentración de tRNA que la mayoría de los genes altamente expresados (GAEs), entonces la premisa fundamental que los predice como de alta expresión —la supuesta mejor correlación entre el uso de codones de GAE y el nivel de tRNA— ya no parece tan robusta. Estos resultados sugieren que el UC por sí mismo no es un buen indicador de niveles de expresión. Desde el punto de vista del genoma, hay muchos genes en *E. coli* K12 que muestran un CRI similar o más alto que las PRs y los GAE (Figura 2.3), pero esto no significa necesariamente que también son altamente expresados. En vez de esto sólo se puede proponer con confianza que son traducidos eficientemente, independientemente del número de mRNAs y/o proteínas producido.

2.7 El uso de aminoácidos está relacionado con la traducibilidad pero no es el factor de mayor impacto en las frecuencias de codones.

Desde la perspectiva de las metodologías actuales, que asumen una nula o mínima influencia del uso de aminoácidos (UAA) en la traducibilidad de los genes (Capítulo I, Sección 1.5; Capítulo II, Secciones 2.2 y 2.6.2), podría argumentarse que el CRI tiene un defecto en su diseño al no introducir una corrección en las frecuencias de codones para compensar el efecto de la composición de aminoácidos.

Tabla 2.5. El uso de aminoácidos (UAA) y su relación con el tRNA a diferentes velocidades de crecimiento (columna 1). Se puede apreciar que el UAA genómico (columna 2) correlacionas mejor con la abundancia de tRNA que el conjunto de mRNA transcritos (Tabla 2.2) (columna 3) y las proteínas ribosomales (PRs) (columna 4).

Tasa de crecimiento (duplicaciones / hora)	Correlación† entre la concentración de tRNA y:		
	UAA genómico	UAA en el mRNA	UAA en PRs
0.4	0.8059	0.7476	0.6721
0.7	0.8265	0.7624	0.6823
1.07	0.8421	0.7558	0.6673
1.6	0.8218	0.7566	0.6957
2.5	0.8254	0.7645	0.7128

† Se aplicó el factor de correlación de Pearson, todas las correlaciones son significativas ($p<0.01$). Las diferencias de correlación entre cualquier par de columnas también son significativas (*t-test* para muestras pareadas; $p < 0.001$).

La Tabla 2.5 muestra la correlación entre las concentraciones de tRNA y el UAA en (1) todo el genoma, (2) las PRs y (3) los mRNAs transcritos en cada condición de crecimiento. Se puede apreciar que el UAA y el UC genómico (ver también Tabla 2.4) mantienen una correlación muy similar con las abundancias de tRNA, mientras que las correlación UAA–tRNA de las PRs y el mRNA producido caen significativamente. Para cada uno de los tres grupos de genes analizados, las correlaciones se calcularon de la siguiente manera: por cada aminoácido la suma de las concentraciones de todos los tRNAs que lo traducen fue correlacionada con la suma de las frecuencias de todos los codones sinónimos que lo codifican.

La Figura 2.6 muestra como el UAA genómico tiene una menor dispersión con respecto al contenido de tRNA que el UAA en las PRs y el mRNA transcrito, en cada condición de crecimiento. La caída de la correlación en el mRNA y PRs puede explicarse, en parte, por las restricciones estructurales y funcionales asociadas con el uso extensivo de aminoácidos básicos (e.g. lisina) involucrados en las interacciones mRNA–proteína [163], y a la relativa baja abundancia del único tRNA disponible para traducir los codones de lisina (ver Tabla 2.1). Estos requerimientos estructurales son aparentemente compensados por la síntesis de grandes cantidades de mRNA y el uso preferencial del codón óptimo para lisina. Sin embargo esta tendencia no es perfecta, dado que se sabe que las PRs usan codones raros [164].

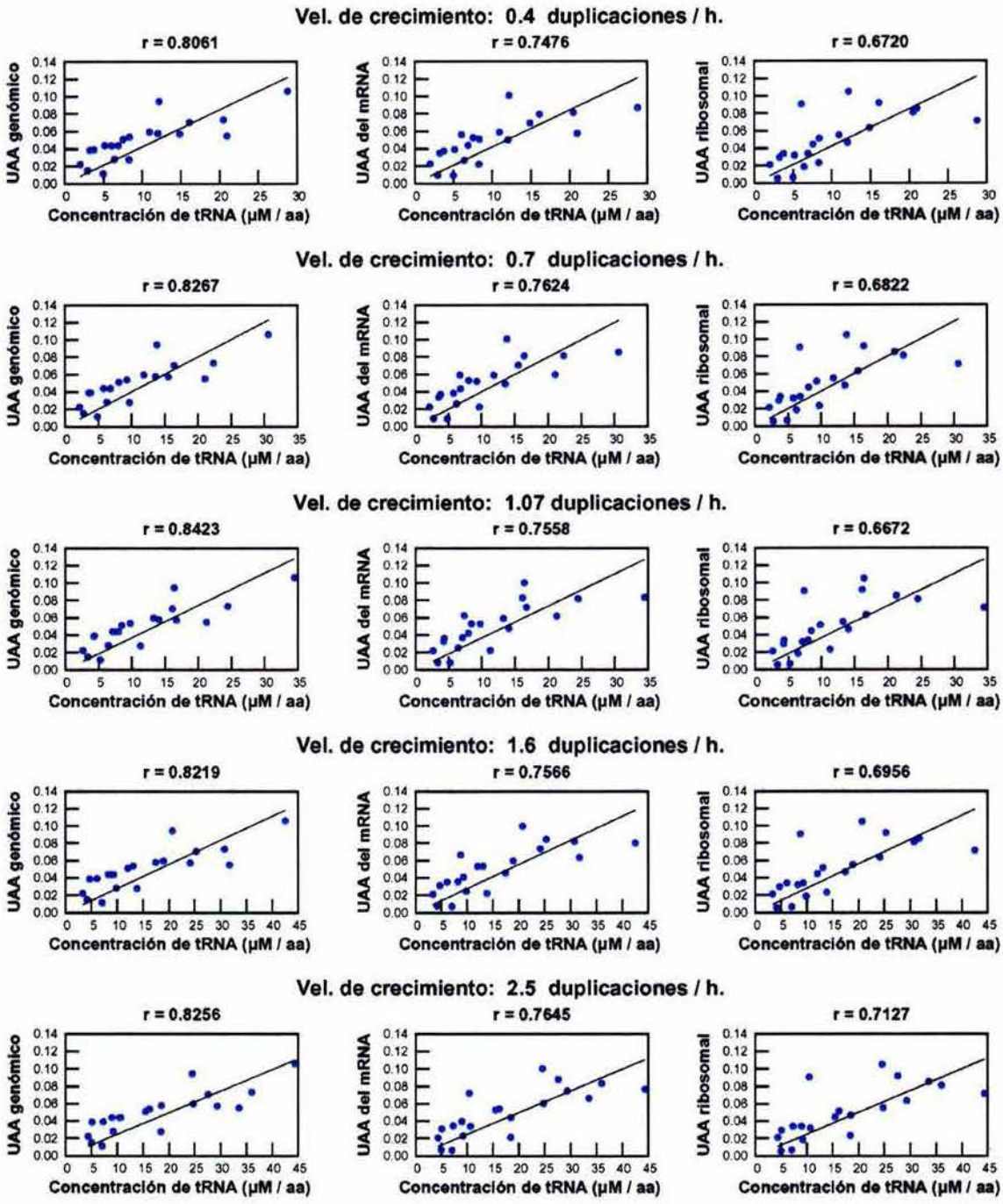

Figura 2.6 Covariación entre la concentración de tRNA en *E. coli* K12 (eje de las Xs) y el uso de aminoácidos (UAA) (eje de las Ys) a través de 5 condiciones de crecimiento. Cada columna de paneles corresponde a un grupo de genes: todo el genoma (columna 1); el mRNA sintetizado (columna 2) y las proteínas ribsomales (PRs) (columna 3). Los renglones de paneles corresponden a correlaciones estimadas por cada una de las 5 velocidades de crecimiento analizadas. El cálculo de la concentración de tRNA que traduce cada aminoácido se detalla en el texto (Sección 2.7).

Además, como se expone en la Sección 2.8, las PRs están dispersas en el espacio de UC del genoma. La relación UAA–tRNA se ha observado anteriormente [93, 165], sin embargo debe notarse que la correlación es mejor con el UAA genómico que con los grupos de genes individuales aquí manejados, hecho que enfatiza la importancia de tomar en cuenta la contribución de la composición genómica de aminoácidos al estudiar la traducibilidad.

Para ponderar el impacto del UAA en el UC, se diseñaron dos índices adicionales. Primero, el índice de riqueza de aminoácidos (AARI por sus siglas en inglés: *Amino Acid Richness Index*) que cuantifica el grado en que genes individuales utilizan los aminoácidos más abundantes en el genoma. La metodología es muy similar a la del CRI y se describe en la Sección 2.7.1. Segundo, el índice de similitud con las proteínas ribosomales (RLI por sus siglas en ingles *Ribosomal–Like Index*) para cuantificar el grado en que genes individuales utilizan los codones más abundantes en el conjunto de proteínas ribosomales de cada genoma. De nueva cuenta, la metodología es muy similar a la del CRI y se describe en la Sección 2.7.2.

Las correlaciones CRI–AARI y RLI–AARI para *E. coli* K12 son muy bajas pero significativas ($r = 0.4$ y $r = 0.16$ respectivamente; $p<0.01$), indicando que si bien por definición el UAA tiene influencia sobre el CRI y RLI (pues no se normaliza la frecuencia de cada codón por la frecuencia total de todos sus codones sinónimos), claramente no es el factor dominante en su comportamiento. Este hecho también es evidente a partir de la elevada correlación ($r = 0.94$, $p<0.01$) entre el CAI —no influenciado por el UAA— y el RLI. La correlación CRI–CAI ($r = 0.71$, $p<0.01$), es suficientemente elevada como para argumentar que tanto el CRI como el CAI están gobernados principalmente por los mismos factores. Resumiendo, dada las evidencias reportadas a favor de que el UAA está relacionado con la eficiencia de la traducción [93, 94], más los resultados aquí presentados (ver Tabla 2.5 y Figura 2.6), la influencia que ejerce el UAA sobre el UC debe enriquecer al CRI como una medida de traducibilidad, pues contribuye a la elevada correlación CRI–tRNA mostrada en la Sección 2.6.3.

Todo lo visto hasta este momento en el Capítulo II sugiere que las concentraciones de tRNA deben ser tales que la célula pueda expresar adecuadamente todos sus genes en el momento que así lo requiera —independientemente del sesgo composicional de los genomas. Por lo menos en *E. coli*, la relación tRNA–UC–UAA a nivel genómico refleja esta armonía entre el genoma y la maquinaria de la traducción. Con base en este razonamiento, no se considera descabellado

asumir que el UC genómico es una referencia adecuada para medir traducibilidad en otros genomas. Por lo tanto, en el análisis comparativo presentado en la Sección 2.8 se utiliza al CRI como medida de traducibilidad y al RLI como medida de niveles de expresión, por ser este último compatible con el CAI pero más fácil de calcular pues se basa exclusivamente en el conjunto de PRs. No obstante, los resultados deben interpretarse con cuidado porque este supuesto pudiera no cumplirse en otros genomas.

2.7.1 El índice de riqueza de aminoácidos (AARI)

El AARI (por sus siglas en Inglés *Amino Acid Richnes Index*) se calcula como el CRI aplicando la Ecuación 1.1 (Capitulo I, Sección 1.5.1). Sin embargo $p_a(c)$, ahora debe interpretarse como la probabilidad del aminoácido c en el genoma a. Para mayor claridad llamaremos a esta probabilidad $pAA_a(c)$. De igual forma $q_{a,i}(c)$, es ahora la frecuencia relativa del aminoácido c en el gene $G_{a,i}$ y se representará como $qAA_{a,i}(c)$ (se cumple la condición $\sum_{c=1}^{20} pAA_a(c) = 1$). La ecuación para calcular el AARI es entonces:

$$AARI_a(G_{a,i}) = \sum_{c=1}^{20} pAA_a(c) * qAA_{a,i}(c). \qquad (2.1)$$

El AARI fue diseñado para ponderar la influencia del UAA en el UC (ver Sección 2.7).

2.7.2 Índice de similitud con las proteínas ribosomales (RLI)

El RLI (por sus siglas en Inglés *Ribosme–Like Index*) fue creado usando la misma estrategia que se empleó para crear el CRI (Capítulo I, Sección 1.5). El RLI cuantifica el grado en que genes individuales utilizan los codones más abundantes en el conjunto de proteínas ribosomales (PRs) de cada genoma. Por lo tanto, el índice tiene un sesgo intrínseco en su definición, donde los genes con uso de codones (UC) similar a las proteínas ribosomales tendrán valores más altos.

El RLI del gene $G_{a,i}$ con base a las PRs de su propio genoma, a, se calcula de manera similar al CRI aplicando la Ecuación 1.1 (Capítulo I, Sección 1.5.1), excepto que ahora el término $p_a(c)$ es sustituido por $pPR_a(c)$ e interpretado como la probabilidad de encontrar el codón c en el conjunto de PRs del genoma a (se cumple la condición $\sum_{c=1}^{64} pPR_a(c) = 1$).

$$RLI_a(G_{a,i}) = \sum_{c=1}^{64} pPR_a(c) * q_{a,i}(c). \qquad (2.2)$$

La Figura 2.7 muestra el RLI para *E. coli* K12. Como es de esperarse, de manera muy similar a la gráfica del CAI (Sección 2.6.2, Figura 2.4), las proteínas ribosomales (círculos rojos) y los GAE (círculos verdes) tienen valores muy altos. Esto se debe a que la PRs son el conjunto de referencia y que los genes altamente expresados (GAE) se definirían como aquellos mostrando valores elevados de RLI. Como la correlación RLI–CAI es muy elevada ($r = 0.94$, $p < 0.01$) y el CAI no esta influenciado por el uso de aminoácidos, no es sorprendente que la correlación AARI–RLI sea muy baja aunque estadísticamente significativa ($r = 0.16$; $p < 0.01$).

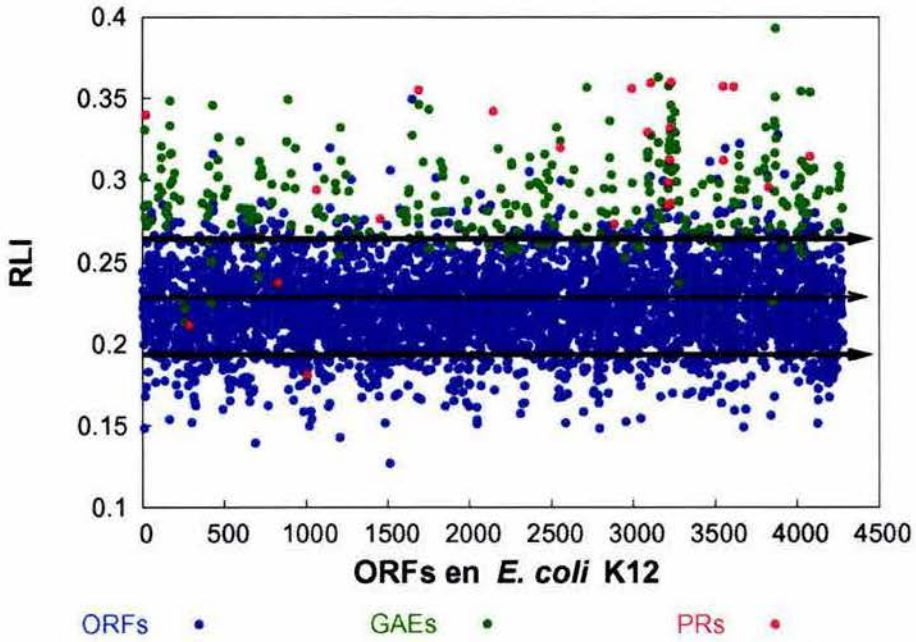

Figura 2.7. El índice de similitud con las proteínas ribosomales (RLI) de *E. coli* K12. El eje de las Xs presenta los genes ($G_{a,i}$) de acuerdo a su orden de aparición en el cromosoma. El eje de las Ys expresa el valor de $RLI_a(G_{a,i})$ para cada gene anotado. La gráfica muestra que las proteínas ribosomales (PRs) y los genes conocidos/predichos como altamente expresados (GAEs) [148] tienen en su mayoría valores altos de RLI. Las flechas indican los distintos niveles de RLI calculados con la misma estrategia que para el CRI (ver Capítulo I, Figura 1.2 y Sección 1.5.2).

2.8 Genes altamente expresados con un uso de codones óptimo en un genoma no tienen las mismas propiedades composicionales en otros genomas.

Esta sección y la siguiente abordan la segunda pregunta planteada en los objetivos (Sección 2.3), sobre la universalidad de los supuestos actuales para predecir niveles de expresión en múltiples genomas donde se conoce poco o nada sobre concentraciones de tRNA. En el organismo modelo *E. coli* K12, un 98% de los GAEs [104, 148] exhiben alto CAI (ver Figura 2.3) y un 90% alto RLI (ver Figura 2.7). Tal distribución es de esperarse dadas las definiciones técnicas de los índices (Secciones 2.6.2 y 2.7.2). Por otro lado, un 92% de los genes altamente expresados (GAE) muestran un CRI típico–alto o rico (ver Figura 2.3). Si los supuestos comunes para predecir GAEs en otros organismos (ver Sección 2.2) están bien fundamentados, entonces el sesgo en UC mostrado por los GAEs en *E. coli* K12 (medido con el CRI y RLI) debe ser similar en otros genomas, independientemente de cuales sean las preferencias de codones características de cada organismo.

Este análisis es necesario debido a que la predicción de GAEs en otros organismos mediante las metodologías actuales, requiere de conocer *a priori* un conjunto de GAEs experimentalmente caracterizados en el genoma(s) de interés —con la excepción de un método reciente que primero busca en el genoma los genes con el mayor sesgo de UC y luego en base a ellos calcula el CAI [162]. Cuando esta información no está disponible, se asume que los genes ortólogos a GAEs en organismos modelo son también altamente expresados y que están optimizados en su UC para ser traducidos con elevada eficiencia. Es decir, que este conjunto hipotético de GAEs también se transcribe abundantemente y que su UC correlaciona óptimamente con las concentraciones de tRNA desconocidas en el genoma bajo estudio. Tal cascada de supuestos implica que un número sustancial de genomas muestran una tendencia común a inducir un fuerte sesgo hacia un UC rico en mas o menos el mismo conjunto de genes (GAEs) que pertenecen a unas pocas categorías funcionales [90, 91, 148]. Dada la cantidad de genomas totalmente secuenciados a la fecha, y a las anotaciones funcionales disponibles, es posible someter a prueba la fuerza de los supuestos mencionados mediante un estudio de genómica comparativa.

Es necesario buscar un conjunto de genes que sea común a todos los genomas y que muestre un CRI alto, para después compararlo con el conjunto de genes con alto RLI que se

obtenga, y determinar si los genes encontrados pertenecen al conjunto característico de categorías funcionales de genes altamente expresados. Con este fin se obtuvieron todos los pares de posibles ortólogos entre 148 genomas de procariotes disponibles a la fecha, utilizando la definición de BDBH para detectar ortología (ver Capítulo 1 Sección 1.7). Cabe destacar que no se realizarán comparaciones numéricas de CRI o RLI entre ortólogos como en el Capítulo I, sólo se determinará si dos genes ortólogos tienen alto CRI y/o alto RLI con respecto a sus propios genomas. Esto se debe a que los valores $CRI_a(G_{a,i})$ y $CRI_b(PO(G_{a,i},b))$ no son directamente comparables por pertenecer a escalas diferentes (revisar notación matemática en el Capítulo I, Secciones 1.5, 1.7 y 1.8)

Tomando un genoma *a* como referencia, se ordenaron todos sus genes $G_{a,i}$ ascendentemente de acuerdo a $CRI_a(G_{a,i})$, luego el CRI de todos los BDBH o probables ortólogos en cualquier otro genoma *b*, $CRI_b(PO(G_{a,i},b))$, fueron colocados en exactamente el mismo orden que su BDBH en el genoma referencia. Al graficar tanto el CRI como el RLI con esta estrategia, es posible observar, en primer lugar, si los genomas tienden a inducir un sesgo similar en el UC de algunos de los genes que comparten; y en segundo lugar, si existe un conjunto de genes con UC óptimo (alto-CRI y/o alto-RLI) común a todos los genomas (ver Sección 2.9).

Como se ilustra en la Figura 2.8, el conjunto de genes con UC rico (alto CRI) es diferente de genoma a genoma. Posiblemente esto se debe a fluctuaciones aleatorias en el CRI ya que la zona del CRI típico parece garantizar tasas de traducción adecuadas —gracias a una buena correlación con la concentración de tRNA (ver Sección 2.5)— y donde la mayor parte de la variabilidad en la secuencia no tiene un efecto significativo en la adecuación de los genes a su contexto genómico. Es decir, las propiedades composicionales del organismo, desde un punto de vista dinámico, es muy robusta contra perturbaciones introducidas por mutaciones. Análisis estadísticos de tripletes en genomas completos ya han llegado a este tipo de conclusiones [166].

Por otro lado, conforme aumenta la distancia filogenética las restricciones nutricionales propias de cada nicho pueden ser tan diferentes que un grupo diferente de genes es "optimizado" para maximizar la eficiencia de la traducción. Por ejemplo, la bacteria *Pseudomonas aeruginosa* puede crecer tan rápido como *E. coli* pero no tiene una sola proteína ribosomal con alto CRI, sin embargo, esta bacteria tiene un número inusual de genes reguladores, transportadores y de

secreción con alto CRI, lo cual podría correlacionar con su extraordinaria versatilidad metabólica y adaptabilidad a ambientes heterogéneos, poniendo así de manifiesto su potencial patogénico.

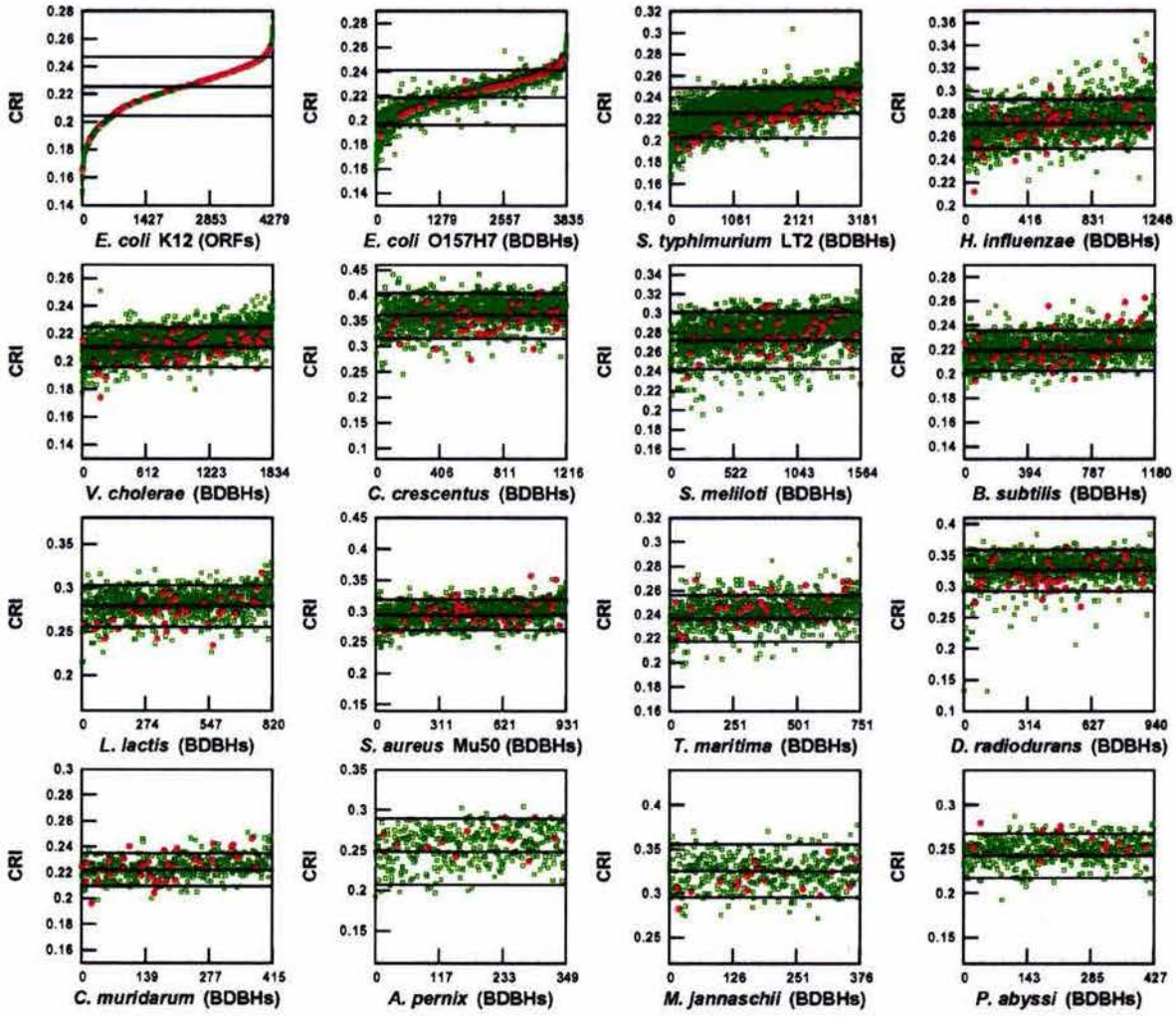

Figura 2.8. Divergencia del uso de codones (UC) a través de múltiples genomas utilizando el índice de riqueza de codones (CRI).. El panel superior izquierdo es el CRI del genoma referencia *a*, en este caso el eje de las Ys grafica los valores $CRI_a(G_{a,i})$, contra el cual que se compara el CRI de los BDBH (ver Capítulo I, Sección 1.7) en los demás genomas *b*. Para todos los demás paneles, el eje de las Ys grafica los valores $CRI_b(PO(G_{a,i},b))$. El eje de las Xs grafica para el panel referencia a los genes $G_{a,i}$ y para el resto de los paneles representa los $PO(G_{a,i},b)$ colocados en exactamente el mismo orden que el $G_{a,i}$ respectivo. Se puede apreciar claramente que los genes con alto CRI son diferentes de genoma a genoma. Las proteínas ribosomales (PRs) se muestran como círculos rojos y todos los demás genes como cuadrados verdes. Las líneas horizontales representan los umbrales de CRI explicados en el Capítulo I, Sección 1.5.2.

No es sorprendente que el análisis utilizando el RLI (ver Figura 2.9) indique que las proteínas ribosomales (PRs) (puntos rojos) tienden a mostrar alto RLI en la mayoría de los genomas, puesto que ellas forman el conjunto de entrenamiento para crear el RLI. Sin embargo, si las PRs son retiradas de la gráfica, el resto de los genes muestran tanta variabilidad en RLI como los genes en la Figura 2.8, a pesar del hecho de que la mayoría de los genes con alto RLI serían predichos como altamente expresados por las metodologías actuales de predicción. En promedio, el 72% de las PRs de todos los genomas muestran un RLI alto, lo cual indica que estas proteínas exhiben en general preferencias similares de codones al interior de cada genoma. Sin embargo, este hecho no involucra que el UC en las PRs y otros genes altamente expresados (GAEs) también mostrará un sesgo hacia un CRI rico, como se puede observar al comparar las Figuras 2.8 y 2.9. La distribución del CRI de las PRs a través de los genomas indica que un 37%, 41% y 14% de las PRs muestran un CRI típico–bajo, típico–alto y alto respectivamente, sugiriendo que los GAEs son en realidad un subconjunto del total de genes eficientemente traducibles y por lo tanto un UC típico constituye efectivamente una zona de tolerancia o seguridad donde los genes pueden alcanzar tasas adecuadas de traducción.

2.9 Con los datos actuales no parece existir un conjunto de genes con UC óptimo que sea común a todos los genomas

Se buscó identificar un conjunto de genes que tuviera alto CRI o alto RLI en todos los genomas. Debido a que el tamaño del genoma tiene un efecto directo en el número de BDBHs (ver Capítulo I, Sección 1.7), sólo se tomaron en cuenta genomas más grandes de 10^6 pb. Si se considera a todos los procariotes no es posible encontrar un solo gene que siempre muestre alto CRI o alto RLI. El mismo resultado se obtuvo cuando el criterio de búsqueda se relajó a genes con valores típico-alto o alto para los dos índices (sólo se encontró un gene). Finalmente la búsqueda se restringió únicamente a eubacterias y el rango aceptado de UC fue expandido a genes con CRI y RLI típico o alto. De un total de 79 genes comunes a las eubacterias, 28 genes mostraron consistentemente un CRI típico o alto y 58 genes un RLI típico o alto. Todos los genes que se encontraron mediante el CRI fueron también encontrados con el RLI. Como es de esperarse, los otros 30 genes restantes encontrados sólo por el RLI son esencialmente proteínas ribosomales (21 PRs, 3 factores de traducción y seis con otras funciones), lo cual es obvio dado que las PRs constituyen el conjunto de entrenamiento del RLI (ver Métodos, Sección 2.7.2).

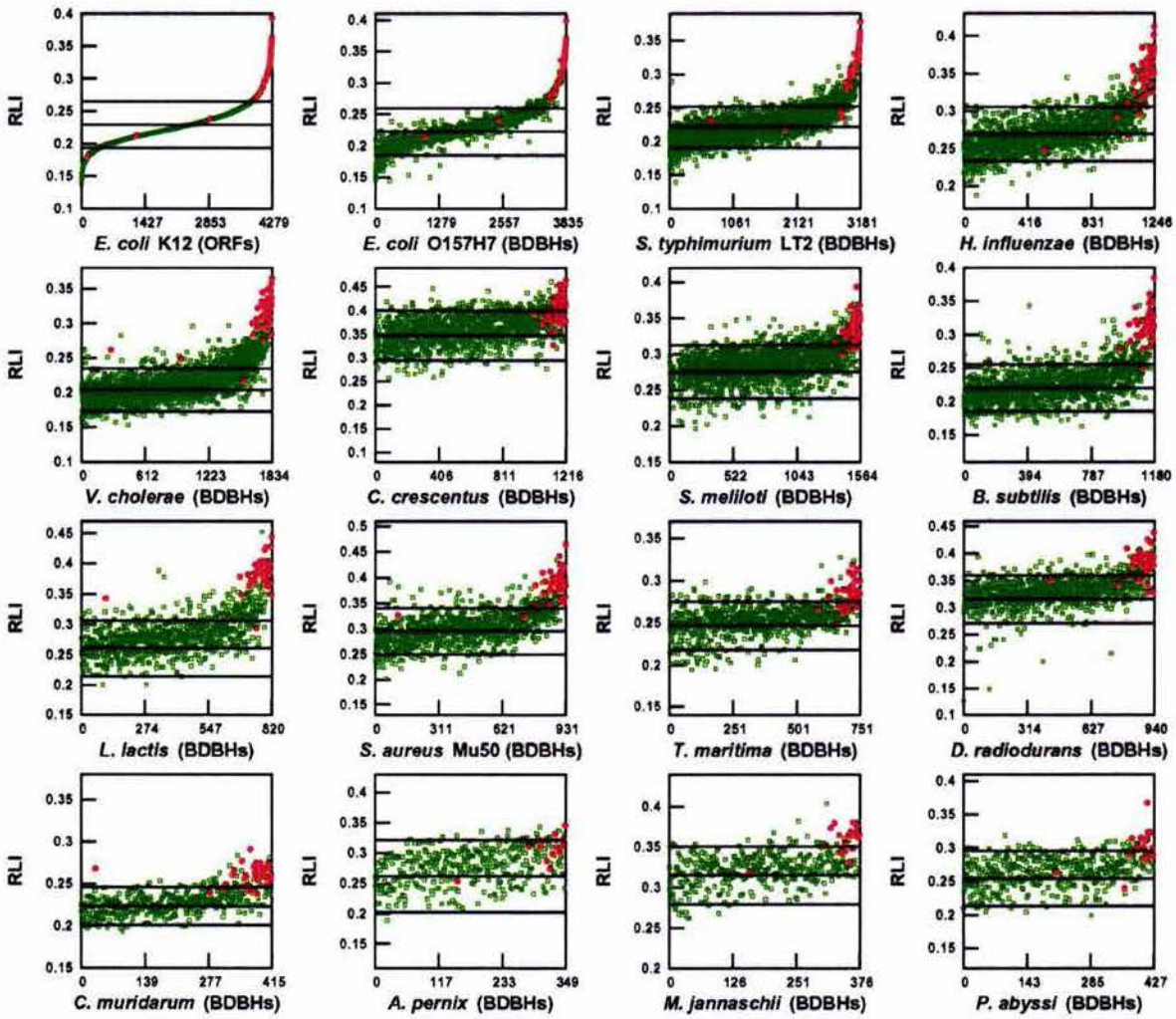

Figura 2.9. Divergencia del uso de codones (UC) a través de múltiples genomas utilizando el índice de similitud con las proteínas ribosomales (RLI). El orden de los BDBHs (ver Capítulo I, Sección 1.7), la ubicación de los genomas en la gráfica y la descripción de los colores es el mismo que en la Figura 2.7. Las líneas horizontales representan los umbrales de RLI explicados en el Capítulo I, Sección 1.5.2.

Si la búsqueda se realiza sólo en firmicutes o proteobacterias, el número de genes comunes aumenta pero la tendencia es la misma, es decir, que un 98% de los genes encontrados por el CRI también son encontrados por el RLI y los genes que no encuentra el CRI están relacionados con la traducción. Por supuesto, el encontrar un conjunto común de genes entre organismos podría indicar que estos genes son importantes, pero también podría significar

meramente que los organismos son parientes evolutivamente cercanos. Estos análisis indican que no es posible encontrar un grupo de genes que muestre un UC rico utilizando como referencia ya sea el UC genómico o el UC en las PRs (a no ser que un UC óptimo se defina como genes mostrando CRI/RLI típico o alto). Los genes con UC rico en un genoma tienen principalmente un UC típico (típico-bajo o típico-alto) en otro. El hecho de que las PRs muestren un alto RLI es únicamente debido al sesgo introducido por la definición del índice. El no poder encontrar un conjunto común de genes con alto RLI, conteniendo al menos algunas PRs, sugiere fuertemente que no hay un conjunto de genes cuyo UC esté universalmente optimizado.

2.10 Discusión

En el Capítulo I se propuso que el nivel típico de UC implica una zona de tolerancia o seguridad donde los genes pueden alcanzar niveles apropiados de expresión. La magnitud de la correlación entre el UC genómico (UC_G) y la concentración de tRNA apoyan fuertemente esta idea (ver Sección 2.5). El beneficio obtenido por los organismos al codificar codones, no necesariamente óptimos, leídos por tRNAs abundantes que traducen aminoácidos también razonablemente abundantes, debe verse reflejado en tasas de traducción más elevadas. Puesto que el nivel de expresión depende en gran medida de la condición particular de crecimiento, un UC típico debe constituir una zona de tolerancia o seguridad donde los genes pueden ser expresados adecuadamente sin comprometer peligrosamente la función. De ser así, los genes pueden sufrir mutaciones neutrales, ventajosas o ligeramente desventajosas mientras el nivel de variabilidad introducido no lleve a los genes a la zona más profunda del UC pobre (bajo CRI). Además, como el CRI se basa en el UC_G para su definición y la correlación UC_G–tRNA es mayor que la correlación PR–tRNA, el CRI constituye un mejor indicador de traducibilidad que otros índices de UC disponibles. Esta aseveración encuentra justificación porque los genes con alto CRI muestran la más alta correlación con la disponibilidad de tRNA en comparación con cualquier conjunto de GAE reportado a la fecha (Sección 2.6).

Por otro lado, el hecho de que el UC en genes con alto CRI correlacione mejor con la concentración tRNA que el UC en GAE es discordante con el supuesto principal detrás de las metodologías actuales de predicción de niveles de expresión —el UC en GAE debe mostrar niveles más altos de correlación con las abundancias de tRNA en comparación con genes de menor expresión. Por lo tanto, la fuerza de la asociación predictiva que se supone existe entre el

UC y los niveles de expresión [30, 104, 148], no puede seguirse considerando como confiable aún en genomas con sesgos pronunciados de UC. Por supuesto, no se pretende negar la existencia de una correlación entre el UC y el nivel de expresión, de hecho las correlaciones mRNA–tRNA y GAE–tRNA en la Figura 2.5 y Tabla 2.4 indican que sí existe y es significativa, pero no es tan elevada como para argumentar que se trate de una asociación predictiva. Solo es posible afirmar con confianza que los GAEs son un subconjunto de todos los genes traducidos eficientemente en el genoma. Es importante recordar que el nivel de expresión también depende fuertemente de otros factores genéticos (e.g. fuerza del promotor, modo de regulación, tasa de producción de mRNA, vida media tanto del mRNA como de la proteína, etc.), lo cual sugiere que aquellos métodos que busquen predecir niveles de expresión deben tomar en cuenta estos factores en sus modelos.

Se sabe que en algunos genomas la deriva génica puede influenciar significativamente el UC y el UAA [151, 152, 167], empero aún así las concentraciones de tRNA son tales que todos los genes se expresan en los niveles apropiados. Por lo tanto, se espera que aún en genomas sin sesgos significativos en sus preferencias de codones, tanto el UC como el UAA correlacionen bien con las abundancias de tRNA. De ser así, independientemente de la composición de codones puede predecirse una relación entre UC y traducibilidad. Para abordar esta hipótesis primero es necesario medir las concentraciones de tRNA en genomas que no tienen sesgos claros en su UC. Estos datos confirman que el UC no está determinado principalmente por presiones selectivas en el nivel de expresión. Por ejemplo, aún cuando *Pseudomonas aeruginosa* puede alcanzar tasas de crecimiento tan elevadas como *E. coli*, se ha visto que el nivel de expresión no es el factor determinante de su sesgo en UC sinónimos [85], de hecho tiene un numero inusualmente alto de PRs con CRI pobre.

Mediante el análisis de 148 organismos cuyos genomas han sido completamente secuenciados, se muestra consistentemente que las PRs y otros GAEs no despliegan un UC óptimo en todos los genomas, de hecho muestran preferentemente un UC típico. Como resultado, no fue posible encontrar un conjunto de genes con UC óptimo común a todos los genomas. Se puede concluir, primero, que es incorrecto asumir que los GAE con un UC óptimo en un organismo modelo, también tendrán un UC óptimo en otros organismos donde no se conoce el nivel de expresión ni las concentraciones de tRNA; y segundo, que se han sobreinterpretado, o incorrectamente generalizado, los resultados de los análisis efectuados en organismos modelo.

Para explicar porque los genomas tienen diferencias tan pronunciadas en su UC, es necesario hacer alusión a múltiples factores: (1) fluctuaciones aleatorias; (2) distancia filogenética; (3) simplemente porque un UC típico garantiza una traducción adecuada, (4) economía de la energía; (5) restricciones nutricionales propias del nicho de cada organismo; y (6) una combinación de estos factores. Los GAEs pueden mostrar en algunos genomas un CRI típico–alto o alto, pero esto no significa necesariamente que todos los genes con UC similar a los GAEs serán también altamente expresados. Sin embargo, si es posible argumentar que pueden ser eficientemente traducidos (dado que despliegan principalmente un UC típico) independientemente de la cantidad de mRNA o proteína producida. Por lo tanto, predecir niveles de expresión a partir únicamente del UC es potencialmente peligroso, involucra argumentos circulares y puede llevar a conclusiones incorrectas.

2.11 Perspectivas

Se tiene contemplado hacer un análisis detallado de los genes que tienen alto CRI en todos los genomas, para determinar si existen evidencia funcionales/selectivas que expliquen por que estos genes tienen alto CRI. Para este propósito es necesario relacionar cuidadosamente la función de los genes con los requerimientos nutricionales propios del nicho de cada organismo.

Se ha visto que los GAEs tienen un sitio de unión al ribosoma (RBS, por sus siglas en inglés *Ribosome Binding Site*) más fuerte que el gene promedio [149]. Dado que los GAEs tienen en su mayoría un CRI típico–alto o alto, sería interesante ver: (1) si esta es una propiedad exclusiva de los GAEs o más general de los genes con CRI típico–alto o alto; y (2) Si hay una tendencia en los genes con CRI típico–alto y alto a tener promotores más fuertes en contraste con genes exhibiendo CRI típico–bajo y bajo.

Anexo I

La ciencia no es "el sentido común organizado"; en el caso más excitante, nos permite reformular nuestra visión del mundo al imponer teorías poderosas contra los muy ancestrales prejuicios antropocéntricos que solemos encubrir bajo el término intuición.

STEPHEN JAY GOULD

Se incluye el artículo tal cual fue publicado en la revista internacional *Molecular Biology and Evolution*, donde se reportan los principales resultados y conclusiones descritos en el Capítulo I. Para evitar redundancia lo más posible, muchos de los argumentos y discusiones en el artículo fueron omitidas en la tesis. Se recomienda fuertemente leer el artículo para tener una idea más completa de todas las evidencias que soportan las hipótesis planteadas.

Successful Lateral Transfer Requires Codon Usage Compatibility Between Foreign Genes and Recipient Genomes

Arturo Medrano-Soto, Gabriel Moreno-Hagelsieb,*[1] Pablo Vinuesa,† J. Andrés Christen,‡ and Julio Collado-Vides**

*Program of Computational Genomics and †Program of Molecular and Microbial Ecology, Centro de Investigación sobre Fijación de Nitrógeno (UNAM), Cuernavaca, Morelos, México; and ‡Department of Probability and Statistics, Centro de Investigación en Matemáticas, Guanajuato, Guanajuato, México

We present evidence supporting the notion that codon usage (CU) compatibility between foreign genes and recipient genomes is an important prerequisite to assess the selective advantage of imported functions, and therefore to increase the fixation probability of horizontal gene transfer (HGT) events. This contrasts with the current tendency in research to predict recent HGTs in prokaryotes by assuming that acquired genes generally display poor CU. By looking at the CU level (poor, typical, or rich) exhibited by putative xenologs still resembling their original CU, we found that most alien genes predominantly present typical CU immediately upon introgression, thereby suggesting that the role of CU amelioration in HGT has been overemphasized. In our strategy, we first scanned a representative set of 103 complete prokaryotic genomes for all pairs of candidate xenologs (exported/imported genes) displaying similar CU. We applied additional filtering criteria, including phylogenetic validations, to enhance the reliability of our predictions. Our approach makes no assumptions about the CU of foreign genes being typical or atypical within the recipient genome, thus providing a novel unbiased framework to study the evolutionary dynamics of HGT.

Introduction

The incessant deluge of completely sequenced genomes has boosted the development of lateral genomics, providing new insights into the nature, prevalence, and evolutionary implications of horizontal gene transfer (HGT) in such critical biological processes as the occupation of new niches and speciation (Berg and Kurland 2002; Gogarten, Doolittle, and Lawrence 2002; Brown 2003). Laterally acquired genes are usually expected to display atypical codon usage (CU) mainly due to two lines of evidence. First, there are biases toward poor CU in genes with plasmid, phage, and transposon related functions (Sharp and Li 1987; Médigue et al. 1991). Second, some chromosomes exhibit distinctive regions with atypical sequence compositions (Lawrence and Ochman 1998; Ochman and Bergthorsson 1998; Kaneko et al. 2000). Given that CU is thought to reflect adaptation of genes to the translational machinery of the host (Ikemura 1981, 1982), and foreign genes have not been exposed to the same evolutionary forces as resident genes, it is commonly assumed that foreign genes should exhibit a poorly adapted codon composition. Thus, atypical CU, G+C content, and/or different oligonucleotide frequencies have been routinely used as critical parameters to predict HGT events (Médigue et al. 1991; Karlin, Mrázek, and Campbell 1998; Lawrence and Ochman 1998; Moszer, Rocha, and Danchin 1999; Garcia-Vallvé, Romeu, and Palau 2000). However, cautious evaluations conclude that current methodologies based on unusual codon abundances and base composition alone are poor indicators of HGT (Koski, Morton, and Golding 2001; Wang 2001) and that comparisons among different approaches generate fewer predictions in common than would be expected by chance (Ragan 2001*b*); nevertheless, methods testing compatible null hypotheses are expected to increase their level of agreement (Lawrence and Ochman 2002; Ragan 2002). Based on these evaluations, the need to develop new quantitative models to interpret data and assess confidence has been stressed (Ragan 2001*a*).

A gene is said to display rich CU if it uses preferentially the most abundant codons within the host genome. Accordingly, poor or atypical CU indicates the preferential use of rare codons, and typical CU reflects a balance between abundant and rare codons. In this work we sought to find out what the actual CU level (poor, typical, or rich) displayed by successfully imported genes was at the moment of acquisition. Thus, it was necessary to collect pairs of imported/exported genes that still conserve the compositional footprint of the donor DNA, without making any a priori assumption about their CU level. The four basic premises in our approach are: (1) candidate xenologous genes (CXGs), or pairs of horizontally exported/imported genes, must bear similar codon composition independently of whether the CU level of acquired genes is poor, typical, or rich in the recipient genomes; (2) CXGs should have similar lengths; (3) they should display the highest global identity values at the protein level, and thus they should be detected by current computational methods to infer orthology; and (4) the phylogenetic relationship between detected CXGs must contradict the hypothesis of vertical inheritance. The first premise was used to obtain an initial list of candidate xenologs, and the other three allowed the removal of potential false positives. To assess the CU of genes, we designed the Codon-Richness Index (CRI), which quantifies the degree to which individual genes use the most abundant codons within a reference genome. Genes showing low, average, and high CRI identify the three CU levels defined above.

We followed the strategy depicted in figure 1 to detect pairs of CXGs. First, we calculated a CU profile of potential xenologs for 103 representative genomes (see

[1] Present address: Department of Biology, Wilfrid Laurier University, Waterloo, Ontario, Canada N2L 3C5.

Key words: horizontal gene transfer, codon usage compatibility, comparative genomics, evolution, Bayesian model.

E-mail: amedrano@cifn.unam.mx.

Mol. Biol. Evol. 21(10):1884–1894. 2004
doi:10.1093/molbev/msh202
Advance Access publication July 7, 2004

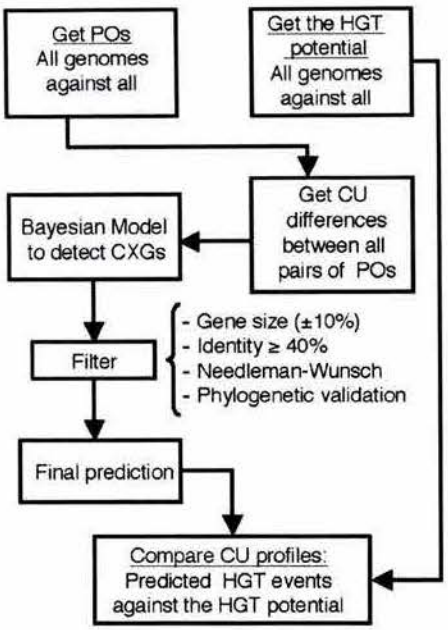

FIG. 1.—Strategy to assess the codon usage (CU) level of genes involved in horizontal transfer events. Putative orthologs (POs), quantification of CU values, candidate xenologous genes (CXGs), and phylogenetic analyses were performed as explained in *Methods*.

table 1 for a complete list), meaning the account of the CU level that each gene from 102 (donor) genomes would display within the remaining 103rd (recipient) genome. Second, we searched for pairs of exported/imported genes that still conserve their original CU, that is, genes that show similar codon frequencies (first working premise). Because CU similarity might not be enough evidence of HGT, we applied additional filtering criteria, including phylogenetic validations, to strengthen the predictions (i.e., the other three working premises; see *Methods* and *Results*). Then, we counted the number of detected xenologous genes at each CU level, from the perspective of their host genomes. Third, we compared the potential HGT CU profile, as obtained from the first step, with the actual HGT CU profile exhibited by predicted xenologs in the second step. The CU comparisons between the actual and the potential CU HGT profiles reveal that most horizontally transferred genes exhibited a typical codon composition at the moment of acquisition, which suggests that poor CU actually represents a strong barrier against successful acquisition and utilization of foreign genes.

Methods

The Codon-Richness Index

To assess and compare the CU level of genes within and across genomes, we designed the CRI based on the genomic overall abundance of all 64 codons. The index quantifies the extent at which individual genes use the most abundant codons within a reference genome. Let $G_{a,i}$ be gene i within genome a, $q_{a,i}(c)$ the relative frequency of codon c in gene $G_{a,i}$, and $p_b(c)$ the probability or relative frequency of codon c in genome b. So, the CRI of gene $G_{a,i}$ based on the codon abundances of genome b is defined as:

$$CRI_b(G_{a,i}) = \sum_{c=1}^{64} p_b(c)*q_{a,i}(c). \quad (1)$$

This index may be interpreted as the expected utility of a particular codon distribution and constitutes a local score function (Bernardo and Smith 1994), where higher values are obtained when gene $G_{a,i}$ uses the most abundant codons within genome b. To obtain the CRI values for genome a relative to its own codon abundances, b must refer to genome a ($b = a$), that is, equation (1) should be evaluated for $CRI_a(G_{a,i})$; figure 2 illustrates the CRI for all genes in *Escherichia coli* K12. By means of this strategy (when $a \neq b$) we may approach the question of whether or not the codon composition of foreign genes is atypical in recipient genomes.

Classification in Three Codon Usage Levels

We classified all genes into three CU categories or levels containing genes with low, typical, and high CRI, as illustrated in figure 2. We initially observed, in *E. coli* K12, that by sorting the genes by their CRI, most genes display approximately a constant CRI difference (see the slope of the curve in fig. 2). Genes with the lowest and highest CRI show greater differences, thus changing the slope of the curve. When similar curves for all genomes were drawn, it became apparent for almost all cases that the inflexion points, where the slope starts to deviate from typical CRI, embrace about 80% of the genes, thereby suggesting a strategy for the definition of genome-specific thresholds for high and low CRI. For each genome, we estimated its CRI histogram and calculated the CRI values (low and high CRI thresholds) at which the interval of maximum density and minimum length contained 80% of the genes.

Detection of Candidate Xenologous Genes

We devised a Bayesian method to compute the posterior probability that two putative orthologs (POs) are CXGs given the CU differences among all related POs and the probability of the null hypothesis (i.e., the chances that they are not xenologs preserving their original sequence composition). The underlying assumption is that CXGs satisfy all criteria of current methods to detect orthology based on protein sequence comparisons. Thus, CXGs were sought across all genomes as POs, which were chosen using the bidirectional best hit (BDBH) working definition of orthology as previously reported (Moreno-Hagelsieb and Collado-Vides 2002*a*). Whenever a BDBH was not detected in a given target genome, we used the top scoring BlastP hit as the PO as long as there was no better hit within the reference genome. This is the Ortholog Higher than Paralog (OHTP) definition (Ermolaeva, White, and Salzberg 2001). All BlastP comparisons were run with a maximum cutoff E-value of 0.001, filtering low information sequences and using the option for a Smith-Waterman final alignment (Schaffer et al. 2001). We assume that CRI

Table 1
The Horizontal Gene Transfer (HGT) Potential Among the Representative Set of 103 Complete Prokaryote Genomes

Genome	Low CRI	Typical CRI	High CRI	Total
A. pernix	220691	73085	14	293790
P. aerophilum	226917	65213	896	293026
S. solfataricus	232056	49256	11342	292654
S. tokodaii	233586	50841	8378	292805
A. fulgidus	267463	25548	200	293211
Halobacterium sp	227533	65423	53	293009
M. thermoautotrophicum	282827	10809	122	293758
M. jannaschii	265417	26490	1947	293854
M. kandleri	216853	76324	763	293940
M. acetivorans	183200	96770	11121	291091
P. furiosus	231181	57219	5106	293506
T. acidophilum	279740	14265	144	294149
T. volcanium	232525	48995	12612	294132
N. equitans	269272	23623	2173	295068
B. longum	192800	94234	6870	293904
C. diphtheriae	113972	162544	16818	293334
C. glutamicum	157905	129792	4941	292638
M. tuberculosis H37Rv	208019	55462	28223	291704
S. coelicolor	262610	24695	118	287423
T. whipplei Twist	158998	101610	34215	294823
A. aeolicus	270363	23211	497	294071
B. thetaiotaomicron VPI-5482	186603	84445	19805	290853
P. gingivalis W83	89249	193897	10576	293722
C. muridarum	200807	54740	39186	294733
C. caviae	196185	55423	43018	294626
C. pneumoniae TW 183	182254	66811	45453	294518
C. tepidum TLS	163277	104917	25190	293384
G. violaceus	199344	65258	26599	291201
Nostoc sp	180834	74293	34375	289502
P. marinus CCMP1375	210574	56508	26667	293749
P. marinus MED4	248298	40397	5224	293919
P. marinus MIT9313	112159	172848	8359	293366
Synechococcus sp WH8102	189138	92556	11420	293114
Synechocystis PCC6803	215477	73496	3491	292464
T. elongatus	197067	88231	7858	293156
D. radiodurans	237132	51092	4276	292500
B. anthracis A2012	228403	52509	9996	290908
B. halodurans	174010	86401	31154	291565
B. subtilis	175368	96220	19931	291519
C. acetobutylicum	256723	28783	6277	291783
C. perfringens	273793	17441	1674	292908
C. tetani E88	273125	18839	1294	293258
E. faecalis V583	222220	61378	8920	292518
L. plantarum	201190	81188	10244	292622
L. lactis	235021	50984	7305	293310
L. innocua	226505	55510	10555	292570
M. gallisepticum	261943	28181	4781	294905
M. genitalium	250284	39796	5067	295147
M. penetrans	279457	14327	810	294594
M. pneumoniae	241071	49230	4641	294942
M. pulmonis	281857	12448	544	294849
O. iheyensis	226221	50551	15359	292131
O. yellows phytoplasma	277629	16775	474	294878
S. aureus Mu50	243684	43046	6153	292883
S. agalactiae 2603	218403	59219	15885	293507
S. mutans	223781	60657	9233	293671
S. pneumoniae R6	193852	81234	18502	293588
S. pyogenes MGAS8232	205537	69398	18851	293786
T. tengcongensis	231937	51944	9162	293043
U. urealyticum	288095	6784	138	295017
F. nucleatum	279680	13578	306	293564
Pirellula sp	103373	156493	28440	288306
A. tumefaciens C58 *UWash*	190657	72356	27216	290229
B. floridanus	277397	14905	2741	295043
B. bronchiseptica	255349	35017	260	290626
B. japonicum	206215	75167	5932	287314
B. melitensis	194500	71907	26029	292436
B. aphidicola	277804	15585	1738	295127
Buchnera sp	276040	17628	1389	295057
C. jejuni	265695	26368	1914	293977
C. crescentus	240000	51026	868	291894
C. violaceum	238489	52425	310	291224
C. burnetii	180682	86053	26887	293622
E. coli K12	149254	135551	6515	291320
G. sulfurreducens	193116	82265	16805	292186
H. ducreyi 35000HP	228811	58743	6360	293914
H. influenzae	236733	49700	7587	294020
H. hepaticus	238135	45260	10361	293756
H. pylori 26695	238771	49350	5946	294067
M. loti	201651	76917	9788	288356
N. meningitidis MC58	154865	138297	390	293552
N. europaea	150040	132992	10138	293170
P. multocida	237421	47818	8377	293616
P. luminescens	164234	98934	27780	290948
P. aeruginosa	251160	36938	1966	290064
P. putida KT2440	208243	74600	7438	290281
P. syringae	185421	82506	22096	290023
R. solanacearum	230872	58669	970	290511
R. palustris CGA009	225993	56191	8630	290814
R. conorii	234721	52704	6832	294257
S. oneidensis	164013	107709	19440	291162
S. meliloti	203487	66631	19310	289428
V. cholerae	133843	152096	5872	291811
V. vulnificus YJ016	143680	136210	10713	290603
W. brevipalpis	289146	5854	20	295020
W. succinogenes	286095	7412	80	293587
X. citri	226500	60461	4358	291319
X. fastidiosa	77286	156537	58976	292799
Y. pestis CO92	102650	181804	6960	291414
B. burgdorferi	263354	28776	1904	294034
L. interrogans	218598	61076	11230	290904
T. pallidum	136104	155162	3369	294635
T. maritima	263902	29201	670	293773
Average	216081.70	66000.35	10678.75	292760.80
Fraction	0.74	0.22	0.04	1.00

NOTE.—Columns 2, 3, and 4 illustrate the number of genes that would enter each reference genome with poor, typical, and high Codon-Richness Index (CRI), if there were an HGT event in this precise moment. Column 5 denotes the total number of potential xenologs.

differences between CXGs follow a normal distribution with mean zero and very small standard deviation, σ_h. Similarly, CRI differences between nonxenologous genes (or POs) were assumed to follow a normal distribution with mean 0 and a standard deviation, σ_o, much greater than σ_h. Hence, the required predictive probability would be

$$P(H(G_{a,i}) = b \mid \bar{D}) = K\frac{\sigma_o}{\sigma_h} e^{\frac{D_{a,i}(b)^2}{2}\left[\frac{1}{\sigma_b^2} - \frac{1}{\sigma_o^2}\right]} P(H(G_{a,i}) = b), \quad (2)$$

where $H(G_{a,i}) = b$ represents the hypothesis that gene $G_{a,i}$ was involved in an HGT event with genome b, that is, that genes $G_{a,i}$ and $G_{b,i}$ are CXGs. $\bar{D}$ is the vector of CRI differences between $G_{a,i}$ and each one of its POs in other

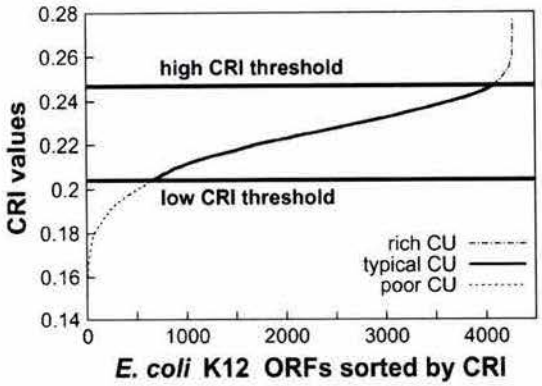

FIG. 2.—The Codon-Richness Index (CRI) of genes sorted in an increasing order. The concepts of poor, typical, and rich codon usage (CU) become apparent. The thresholds for low and high CRI are located at the inflexion points of the curve, embracing about 80% of the genes (see *Methods*).

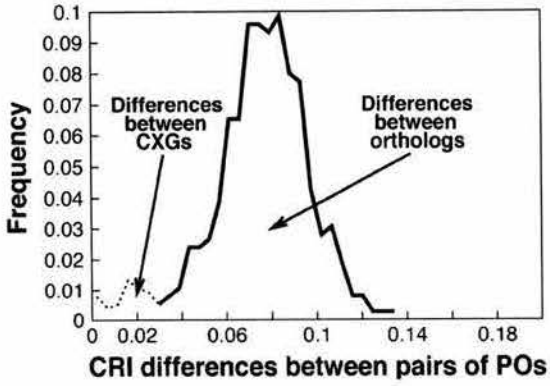

FIG. 3.—Codon-Richness Index (CRI) differences between pairs of putative orthologs (POs) in *H. influenzae* and *N. meningitidis* MC58. The CRI of all pairs of POs in both genomes was calculated based on the codon abundances of *H. influenzae*. Smaller CRI differences can be clearly observed between candidate xenologous genes (CXGs).

genomes (e.g., $G_{b,i}$, $G_{c,i}$, $G_{d,i}$, etc.), based on the codon composition of genome a. $D_{a,i}(b)$ is the specific CRI difference between $G_{a,i}$ and its PO within genome b ($G_{b,i}$), based on genome a. K is the normalization constant. $P(H(G_{a,i}) = b)$ represents the independent prior beliefs we have in the sense that $G_{a,i}$ and $G_{b,i}$ are CXGs; that is, if a reference gene has n POs then we consider a priori that all n POs have equal chances of being the CXGs, and so, taking into account the null hypothesis $H(G_{a,i}) = 0$, that probability is $1/(n+1)$. The threshold to make the initial detection of CXGs was set to the value of the posterior probability of the null hypothesis, that is, evaluating equation (2) for $P(H(G_{a,i}) = 0 \mid \bar{D})$ and $\sigma_h = \sigma_o$. We performed the predictions by stringently defining $\sigma_h = 0.002804$ and $\sigma_o = 0.1$, which translates to the fact that two POs are considered as CXGs if $|D_{a,i}(b)| < 0.0075$ (note that this difference represents in average 1/18 of the CRI scale, see figs. 2 and 3).

Phylogenetic Analysis

It has been demonstrated that the most similar genes in a local or global alignment are not necessarily the closest neighbors in a phylogenetic tree (Koski and Golding 2001). Accordingly, our filtering criteria, where we require that pairs of detected candidate xenologs show small CRI differences and the highest global identity at the protein level, might not be enough supporting evidence. Therefore, for each predicted pair of CXGs we estimated protein-based maximum-likelihood (ML) phylogenies and eliminated all predictions producing trees not consistent with the hypothesis of HGT.

A reference topology to perform phylogenetic incongruity tests was obtained by combining three different strategies and then inferring a consensus tree. First, we used the standard taxonomy for prokaryotic genomes, as reported in the NCBI Taxonomy database (Wheeler et al. 2000). Second, a dendogram was built based on the average protein similarity of all genes shared between each pair of genomes (see elimination of redundant genomes in Moreno-Hagelsieb and Collado-Vides [2002*b*]). A distance matrix was generated with these scores and the program FITCH from the Phylip 3.6b suite of programs (Felsenstein 1989) was used to generate a tree, allowing for overall global rearrangements. The resulting topology using this criterion matched surprisingly well the current taxonomy of prokaryotes, with only five genomes misplaced (data not shown). Third, from the set of genes previously proposed as good candidates to predict evolutionary relationships among prokaryotes (Zeigler 2003), we selected those with at least 90 BDBHs in our set of nonredundant genomes, namely *atpD*, *cysS*, *ffh*, *glyA*, *recA*, and *serS*. Phylogenetic analyses of the products encoded by these genes, as well as for all predicted CXGs, were performed as described below. Only those clades/lineages supported by the three approaches were considered as reference.

Multiple alignments of putative orthologous proteins were performed using the program ClustalW (Thompson et al. 1997) with default settings. Saturated sites and potentially misaligned regions were removed applying the program Gblocks (Castresana 2000) with default parameters. Protein trees were constructed under the ML optimality criterion and applying the JTT+Γ model of amino acid substitutions as implemented in PHYML (Guindon and Gascuel 2003), which accommodates among-site rate variation using a discrete gamma distribution with four rate categories. Nodal support for each JTT+Γ ML phylogeny was assessed by the analysis of 100 bootstrapped alignments and the resulting consensus tree, which were generated with the programs SEQBOOT and CONSENSE, respectively, in the Phylip 3.6b package. Concerning the set of HGT predictions, we initially parsed automatically each tree and regarded it as a good prediction only if the two putative xenologs shared a terminal node, its bootstrap support was $\geq 75\%$, and if a third PO was present such that it rendered the association between the pairs of predicted xenologs phylogenetically incongruent. Finally, all trees that successfully passed such filters were visually inspected for the quality of the topology and potential problems such as the presence of long branches, poorly resolved clades, multiple paralogs, etc. All doubtful predictions were subsequently discarded.

Results

Most Potential Foreign Genes Exhibit Poor Codon Usage from the Perspective of Potential Recipient Genomes

We calculated the CU level of genes by applying a genome-based CU measure, the CRI, designed to quantify the degree to which individual genes use the most abundant codons within a reference genome (see *Methods*). Then, we set two genome-specific CRI thresholds to classify the genes into three classes—the high-CRI (rich CU class), the typical-CRI (typical CU class), and the low-CRI (poor CU class) (see *Methods* and fig. 2). Next, we performed a massive in silico HGT, which consists in considering each genome in turn as recipient and all other genomes as potential donors. We then computed the CRI that each gene (or potential xenolog) from the donor genomes would display within the putative recipient genome. Finally, we counted the number of potential xenologs that fell into each of the three CU categories based on the CRI they would display in the recipient genomes (see *Methods* for details). The CU profile formed by the number of potential xenologs entering a genome as poor, typical, and rich is what we call the potential CU HGT profile.

As shown in table 1, 74% of the genes in other genomes are currently low-CRI genes with respect to the potential recipient genomes, 22% would qualify as typical-CRI genes, and 4% as high-CRI genes. Interestingly, a couple of genomes accept as typical genes most genes in other genomes (see *X. fastidiosa* and *P. gingivalis* W83). If we restrict the HGT potential estimation to genomes belonging to the same taxonomic category, for instance only within the Proteobacteria, a similar pattern is observed, in that most potential xenologs ($\geq$65%) would display poor CU in the recipient genome (data not shown). As expected, the number of potential xenologs with typical CU increases as the evolutionary distances decrease. It is worth noting that the HGT potential is consistent with the expectations and results of several authors who have based their work on atypical gene content (Karlin, Mrázek, and Campbell 1998; Lawrence and Ochman 1998; Garcia-Vallvé, Romeu, and Palau 2000). For instance, from the genes predicted by Lawrence and Ochman (1998) as horizontal acquisitions in *E. coli* K12 strain MG1655, 503 (84%) had a CRI lower than the genomic mean and 329 (55%) are low-CRI genes (we could only find 601 genes identified by name or position out of their 755 predictions; missing genes might have been removed from the current genome version as a result of over-annotations clean up). Similarly, 310 (82%) of the horizontal acquisitions in *E. coli* predicted by Garcia-Vallvé, Romeu, and Palau (2000) are genes below the genomic CRI mean, while 207 (55%) are low-CRI genes (we could only find 376 genes by their name, b-number, or position out of their 381 predictions, probably for similar reasons as before). However, we must emphasize that these numbers (the potential CU HGT profile) are quite different from the CRI distribution of detected HGT events (the actual CU HGT profile) as shown below.

Pairs of Xenologs that Resemble Their Original Codon Usage Can Be Detected as Homologs with Very Similar Codon Usage

To work only with representative genomes, we reduced the 148 available prokaryotic genomes to a non-redundant set of 103, by following a previously reported methodology based on the average protein similarity of shared genes between pairs of genomes (Moreno-Hagelsieb and Collado-Vides 2002*b*). We are aware that the current sample of completely sequenced genomes is not representative enough to ensure we can detect exact pairs of genomes that have been involved in HGT events. Thus, whenever we say we detect a pair of genes exchanged between two genomes, the recipient/donor genomes could well be close relatives of the actual genomes involved. Candidate xenologous genes (CXGs) across all genomes were identified among POs as described in *Methods*. The rationale is that CXGs are a subset of all genes detectable by current methods to identify orthologs. We are interested in HGT events where the transferred genes still resemble their original CU. Thus, we extracted an initial set of potential CXGs from the set of POs by searching for pairs of POs whose CRI difference is close to zero, when both CRIs are computed using the codon composition of either of the two genomes involved. In other words, only those pairs of POs that use to the same extent the most abundant codons within the donor and/or recipient genome are taken into consideration in order to predict CXGs. In figure 3 we show an example, using *H. influenzae* and *N. meningitidis* MC58, to illustrate that CRI differences between CXGs tend to be much smaller than between POs. As would be expected, the number of CXGs showing small CRI differences increases between closely related organisms. We apply a Bayesian method to perform an initial identification of CXGs. The method calculates the posterior probability that two POs with a small CRI difference are CXGs given the CRI differences between all other related POs and the null hypothesis that none of the POs are CXGs (see *Methods*).

However, it is not the CU criterion alone that discriminates candidate xenologs from POs, but the simultaneous application of other filtering criteria as detailed below. The role of the CU similarity criterion is to guarantee that predicted xenologs will have similar CU, which we regard as an obligatory attribute of the type of HGT events we are interested in. The reliability of the predictions is enhanced by the following criteria: First, predicted xenologs must have approximately the same length ($\pm$10%). Second, the global identity, at the protein level, between predicted xenologs must be $\geq$40% (results with higher identity thresholds are shown below). Third, predicted xenologous genes must be the best hits when all related POs are globally compared with the Needleman-Wunsch algorithm (Needleman and Wunsch 1970), using default parameters as implemented within the EMBOSS package (Olson 2002). Fourth, predicted xenologs must be the closest neighbors in a phylogenetic tree, and the topology of the tree must contradict the reference phylogeny of the genomes analyzed (see *Methods* for a detailed explanation). To minimize ambiguous interpretations, we

Table 2
Predicted Number of Genes Involved in Horizontal Gene Transfer (HGT)

Genome	Low CRI	Typical CRI	High CRI	Total
S. solfataricus	0	1	0	1
S. tokodaii	2	0	0	2
A. fulgidus	0	1	0	1
M. thermoautotrophicum	0	1	0	1
M. acetivorans	0	4	0	4
P. furiosus	0	4	0	4
T. volcanium	1	1	0	2
B. longum	0	1	0	1
C. diphtheriae	0	4	0	4
C. glutamicum	0	3	0	3
M. tuberculosis H37Rv	0	3	0	3
S. coelicolor	0	11	0	11
B. thetaiotaomicron VPI-5482	0	7	0	7
C. caviae	0	0	1	1
C. tepidum. TLS	0	8	1	9
G. violaceus	0	4	0	4
Nostoc. Sp	0	2	1	3
P. marinus MIT9313	0	1	0	1
Synechococcus sp WH8102	0	1	0	1
Synechocystis PCC6803	0	2	0	2
T. elongatus	8	1	0	9
D. radiodurans	2	1	0	3
B. anthracis A2012	0	5	0	5
B. halodurans	0	4	0	4
B. subtilis	0	3	0	3
C. acetobutylicum	3	9	0	12
C. perfringens	0	9	0	9
C. tetani E88	4	3	0	7
E. faecalis V583	0	12	0	12
L. plantarum	0	5	0	5
L. lactis	1	2	2	5
L. innocua	0	4	1	5
M. penetrans	0	4	0	4
O. iheyensis	0	4	0	4
S. aureus Mu50	2	0	0	2
S. agalactiae 2603	0	6	1	7
S. mutans	0	2	0	2
S. pneumoniae R6	0	14	0	14
T. tengcongensis	0	2	0	2
F. nucleatum	5	6	0	11
Pirellula. Sp	0	6	1	7
A. tumefaciens C58 UWash	3	30	1	34
B. bronchiseptica	3	19	1	23
B. japonicum	2	29	1	32
B. melitensis	2	20	0	22
C. jejuni	1	2	0	3
C. crescentus	0	10	0	10
C. violaceum	3	18	0	21
C. burnetii	0	0	1	1
E. coli K12	1	20	0	21
G. sulfurreducens	1	8	0	9
H. influenzae	3	2	0	5
H. hepaticus	0	1	0	1
M. loti	4	57	2	63
N. meningitidis MC58	3	7	0	10
N. europaea	4	11	3	18
P. multocida	0	3	0	3
P. luminescens	2	2	0	4
P. aeruginosa	3	22	3	28
P. putida KT2440	1	13	0	14
P. syringae	2	23	0	25
R. solanacearum	3	25	2	30
R. palustris CGA009	0	8	1	9
S. oneidensis	0	15	0	15
S. meliloti	3	49	7	59
V. cholerae	8	7	0	15
V. vulnificus YJ016	1	8	1	10
X. citri	0	21	0	21
X. fastidiosa	0	2	2	4
Y. pestis CO92	0	13	0	13
L. interrogans	0	3	0	3
T. pallidum	0	1	0	1
T. maritima	0	5	1	6
Total	81	615	34	730
Fraction	0.11	0.84	0.05	1.0

NOTE.—Columns 2, 3, and 4 illustrate the number of genes predicted as involved in HGT events for each genome per Codon-Richness Index (CRI) level. Column 5 denotes the total predictions per genome. For convenience only genomes with at least one prediction are included.

only considered predictions involving at least five POs (the pair of predicted xenologs plus three other POs), as previously suggested (Syvanen 1994). The general strategy for HGT detection is summarized in figure 1.

From the analysis of 103 nonredundant genomes, we detected a total of 375 HGT events involving 730 genes (see table 2); some genes are involved in two or more events. Table S1 in the online Supplementary Material gives details on pairs of xenologous genes and their annotated function. About 36% of the predictions involve hypothetical, putative, or unknown proteins, 28% are enzymes (i.e., reductases, transferases, kinases, dehydrogenases, methylases, mutases, synthases), 19% are involved in transport (if putative genes are included, then the number is 27%), 11% are involved in transcriptional regulation, 4% are genes related to mobile elements, and 2% are drug resistance genes.

To assess the level of conservation of predicted xenologs among closely related genomes, we took the 21 *E. coli* K12 genes involved in HGT events (see table 2) and observed how many of them are present in the other three sequenced *E. coli* strains (0157H7, 0157H7 EDL933, and CFT073). Using the BDBH definition of orthology, we found two genes confined exclusively to strain K12 or two strains (K12 and other), six shared by K12 and two other strains, and 13 genes shared by all strains. This is not surprising, as the number of genes shared between closely related genomes is huge. Furthermore, 12 out of the 21 *E. coli* K12 genes involved in HGT have at least two homologs in the other three strains. There are possible explanations for this; foreign genes may coexist with their native homologs and/or, as suggested by other authors, duplication of foreign genes is effectively more common than duplication of indigenous genes (Hooper and Berg 2003). As expected, the number of conserved xenologs with BDBHs elsewhere decreases with increasing phylogenetic distance. For example, none of the 21 genes in *E. coli* K12 predicted to be involved in HGT events has a BDBH in *H. influenzae*, whereas only one gene has a BDBH in *X. fastidiosa*. Unfortunately, the lack of closely related sequenced genomes on both sides of the probable transferences makes it hard to attain reliable conclusions from this analysis.

Most Horizontal Gene Transfers Involve Genes with Typical and Rich Codon Usage

As shown in table 2, most HGT predictions involve typical-CRI genes (84%), with apparently little contribution of genes displaying low or high CRI. This contrasts with both the potential shown in table 1 and the common underlying assumption that most HGTs involve genes displaying predominantly poor CU in the recipient genome at the moment of acquisition. Table 2 does not specify which genes are imported or exported in the predictions, and thus about 365 genes (50%) are expected to be acquisitions. Even if we assume that the 81 HGTs involving low-CRI genes (11% of the total HGT predictions) are all gene acquisitions, we would still have 284 genes (the remaining 39% of imported genes) displaying typical to high CRI. That is, at least 78% (284/365) of all predicted gene acquisitions display typical to rich CU. Furthermore, the overall tendencies observed in table 2 are not significantly affected if we vary the stringency on the minimal identity threshold required between predicted xenologs to detect an HGT event (see table 3).

Our conclusions rely on the assumption that pairs of candidate xenologs satisfying the four filtering conditions can be regarded as evidence of xenologous genes that still resemble their original CU. These conditions are: (1) CXGs must display similar CRI; (2) they must display similar length; (3) they must show the highest global protein identity; and (4) the relationship between CXGs must contradict the hypothesis of vertical inheritance. However, it could be argued that two genes with rather different CU might yield the same CRI score because it is a weighted average, and, consequently, similar CRIs do not necessarily indicate recent HGT events. Although such a scenario is mathematically possible, the combination of the four criteria makes it unlikely to be the case for CXGs. To corroborate this, we took all predicted xenologs showing at least 80% global protein identity and performed codon-wise alignments with their DNA sequences. If CU similarity decays quickly, then most aligned codons should be different. Even for the case of 80% identity, we observe long stretches of DNA sequence identity, which warrants that a substantial fraction of the aligned codons are the same. More specifically, predicted xenologs showing a global protein identity greater than 90% had an average 80% of identical codons in the alignment and 85.6% of average nucleotide identities when running the Smith-Waterman algorithm from the EMBOSS package with default settings. Similarly, those CXGs with global protein identity between 80% and 90% had on average 53% of identical codons and 77% of identities with Smith-Waterman. CXGs with 70%–80% of global protein identity showed an average 72% of DNA sequence identity with Smith-Waterman. Consequently, it is apparent that genes with protein identities $\geq$70% have similar CU vectors, leaving little room, if any, to argue that they are different but had a similar CRI by chance.

The CRI of a gene is most sensitive to codon frequency changes when they affect the frequencies of the most abundant codons in the reference genome. Therefore, if changes occur in codons that contribute little to the score (i.e., rare codons), then the CRI will diverge at a slower rate than the DNA sequence identity. For predictions below 70% sequence identity, more differences in the codon usage vectors are observed, but they are located mostly on codons that have no significant influence on CRI. There are also cases where the most abundant codons are essentially the same for two genomes. In such scenarios, the CRI in both donor and acquired genes will vary even more slowly (e.g., *E. coli* K12 and *N. europaea*). On the other hand, genes exchanged between genomes with large differences in CU would display very poor CRI values (e.g., an HGT from *M. loti* to *H. influenzae*).

Table 3
Horizontal Gene Transfer (HGT) Predictions if the Identity Threshold for Candidate Xenologous Genes Is Gradually Increased

Identity Threshold (%)	Genes Predicted as Involved in HGT Events						
	Low CRI		Typical CRI		High CRI		
	n[a]	%	n	%	n	%	Total
40	81	11	615	84	34	5	730
50	63	10	517	85	28	5	608
60	45	11	328	83	22	6	395
70	31	16	145	76	14	7	190
80	9	17	39	74	5	9	53
90	4	22	12	67	2	11	18

NOTE.—The number of genes predicted as involved in HGT events with low, typical, and high Codon-Richness Index (CRI) when the identity threshold, at the amino acid level, to consider an HGT event as recent is varied from 40% to 90%.

[a] Number of genes predicted as involved in HGT events per codon usage class.

For a better assessment of the role of HGT within each CU class (c), the total number of predicted HGT events per CU class (N_c) should be normalized. One possibility is to divide N_c by the number of comparisons necessary to detect HGTs in the CU class c, that is, the product of the number of resident genes within class c, r_c, and the number of potential foreign genes that would enter directly in the same class (f_c), as indicated by the HGT potential (see table 1), more precisely, $N_c/(r_c f_c)$. Though correct, this normalization would favor our interpretation that very few successful HGTs involve genes with poor CU, since the number of potential foreign genes that would arrive with low CRI is enormous (see table 1). Alternatively, we may relax the normalization criterion and divide N_c only by the number of resident genes in the corresponding class (N_c/r_c). Such normalization favors low-CRI genes, since, by definition, the typical-CRI area contains 80% of the genes (see *Methods*), and thus the number of HGTs entering with typical CRI is diluted. Despite such dilution, HGTs with typical CRI were found to occur 1.2 times more frequently than HGTs involving low-CRI genes, whereas HGTs with low CRI were 1.6 times more frequent than HGT with high CRI. However, if we use the former normalization criterion $N_c/(r_c f_c)$, then for each detected HGT event involving a low-CRI gene there would be 3.7 and 12.7 HGT events involving genes with typical and high CRI, respectively.

One source of potential bias in our results is derived from our phylogenetic congruency tests. If within our nonredundant set of genomes poor-CU genes tend to have

less POs than genes with typical CU, then our predictions might be biased toward genes with typical CU; in our approach less than five POs would exclude such transfers from consideration. Before performing the phylogenetic analyses, there were certainly more HGT predictions because the only filtering criterion was that candidate xenologs had to be the most similar at the protein level; however, even in such circumstances the proportion of genes with poor, typical, and rich CU was essentially the same as that reported in table 2. Although the number of POs will increase as more genomes are sequenced, the proportions of xenologs with poor CU is so low that it is unlikely that it will ever be higher than the proportion of genes with typical CU. Another source of bias might exist between genomes with similar CU; in such cases genes with similar CRI might be found by chance, and the criterion of codon similarity alone could not discriminate between xenologous and orthologous genes. Independently of whether or not the genomes have similar CU, recently exchanged pairs of genes will always display similar CU immediately upon introgression; this situation clearly illustrates and stresses the importance of the phylogenetic and global protein similarity analyses as additional CU-independent filters.

Current methods based on atypical sequence characteristics also have important sources of bias. For example, if a given Open Reading Frame (ORF) is highly atypical it might not actually be a true gene. This seems to be the case for a number of reported HGT predictions (Lawrence and Ochman 1998) that have been eliminated from the current version of the *E. coli* K12 genome in the Entrez genome database. In addition, there are alternative phenomena, besides HGT, that might explain the low GC content and/or poor CU in some genes. For instance, a substantial number of genomes have a GC distribution that is skewed toward low GC, and it was suggested that a remote origin for genes showing low GC seems unlikely. Selection for low GC is a more parsimonious explanation, because functional scenarios involving replication or recombination are easily conceivable for these genes (Syvanen 1994). Genes with low GC might also result from structural constraints, as is the case for ribosomal proteins displaying an excess of lysine residues, coded preferentially by the AAA codon required for RNA protein interactions (Lawrence and Ochman 1997; Ramakrishnan and White 1998). This also explains why most ribosomal proteins do not show high CRI in *E. coli* K12 (AAA is not an abundant codon).

Common Predictions with Previous Reports of Horizontal Gene Transfer

Among the predictions obtained in this work there are some that agree with previous HGT reports. For example, there is strong evidence that *sodC* and *bioC* were transferred between *H. influenzae* and *N. meningitidis* (Kroll et al. 1998). We correctly detect the transference of *bioC*. However, *sodC* is not present in the genome of *H. influenzae* strain Rd as reported in GenBank. Similarly, a substantial fraction of predicted CXGs is involved in transport (19%), regulation (11%), drug resistance, and mobile elements (6%). These functions are often regarded as exchangeable (Gray and Garey 2001; Scott 2002; Beaber, Hochhut, and Waldor 2004).

Given the nature of our methodology, we have very few predictions in common with methods based on atypical sequence characteristics. This is expected, as our method is not designed to be exhaustive and the other methods favor genes with poor CU by definition.

The Case of Horizontal Gene Transfer Between Archaea and *Thermotoga maritima*

We sought another source of information that might either contradict or support the present interpretation. It has been suggested that the eubacterium *Thermotoga maritima* aquired 24% of its genes from an archaeal source (Nelson et al. 1999; Nesbo et al. 2001). Even though this number might be an overestimation (Kyrpides and Olsen 1999; Koski and Golding 2001), the six predictions we obtained involving *T. maritima* genes are all exchanges with Archaea (see table S1 in the online Supplementary Material), five showing typical CU and one showing rich CU (see table 2). The six genes in Archaea currently show typical or rich CU when seen from within *T. maritima*. Furthermore, some archaeal genomes (potential donors) show reasonably good CU compatibility with *T. maritima*. For instance, 64%, 60%, and 58% of all genes in *A. fulgidus*, *P. abyssi*, and *M. jannaschii*, respectively, would show typical to rich CU in *T. maritma* if they were transferred in this moment, despite the divergence that these genomes have suffered since the original speciation events. These results strengthen the idea that genes showing compatible CU with a potential recipient genome are more likely to be successfully transferred. The reader should be aware that we predict very few HGTs due to five reasons: first, we did not attempt to detect every possible xenolog, just those relevant to answer the fundamental question in this analysis, namely what the actual CU level predominantly shown by foreign genes is at the moment of introgression; second, we only take into consideration completely sequenced genomes; third, taxonomic groups are not evenly represented in the set of complete genomes; fourth, detected xenologs are required to share a terminal node in protein trees; and fifth, there must be at least five POs to uphold our predictions.

Poor Codon Usage Represents a Barrier for Horizontal Gene Transfer

Our results imply that foreign genes arriving with rich or typical CU face selection mainly at the functional level, whereas most genes entering the cell with poor CU would most likely be lost in the same way as pseudogenes are eroded, since their functions might not be fully available to assess any functional advantage due to poor translation. Thus, even though the great majority of genes may actually arrive with poor CU, as illustrated by the HGT potential (see table 1), the cell filters them out. Although it is not clear that barriers, beyond the restriction-modification systems, have evolved to prevent lateral gene exchange among prokaryotes (Gogarten, Doolittle, and Lawrence 2002), poor CU might well represent such a barrier as a side effect

of defective translatability. Our interpretation makes biological sense, as integration of foreign DNA into the chromosome is a process subject to efficient surveillance and suppression. Studies of homologous recombination in bacteria show that the frequency of integration of exogenous DNA in the chromosome decreases exponentially as sequence divergence increases (Martin 1999; Denamur et al. 2000; Majewski 2001). In addition, mechanisms preventing illegitimate recombination in prokaryotes and eukaryotes have been reported (Hanada et al. 1997; Wu, Karow, and Hickson 1999). Under the assumption that most HGT events are not expected to be adaptive, but neutral or nearly neutral, the probability of fixation in large populations has been shown to be negligible, most likely leading to the ablation of foreign genes (Berg and Kurland 2002). This is in agreement with our observation that most laterally acquired genes potentially arrive displaying poor CU, but apparently they are strongly selected against due to poor translatability. In contrast, those genes arriving with typical CU have more opportunities to be successfully used by the recipient genome, and thus to persist as long as they provide a reasonable high selection coefficient. We propose that typical CU may well represent a safety or tolerance zone for genes to achieve adequate translation rates and expression levels. This notion is supported by our observation that, at least for *E. coli* K12, genomic codon frequencies (the reference values to calculate CRI) display correlation levels with tRNA concentrations similar to those exhibited by codon abundances in ribosomal proteins (unpublished material). As most highly expressed genes show typical to high CRI, it seems that displaying typical CU is sufficient to attain adequate translation rates.

Our results are also consistent with the work of Smith and Eyre-Walker (2001) who, based on the fact that highly expressed genes in *E. coli* possess rare codons, posit that selection toward codon optimization is a relatively weak force. If a foreign gene, successfully integrated into the genome, displays an acceptable codon composition and provides an adequate amount of protein to perform satisfactorily its function (neutral, nearly neutral, or highly adaptive), then there is no need for an additional strong selective pressure to turn its rare codons into abundant codons.

Codon Usage Amelioration Is Unnecessary

We need to re-evaluate the notion that a foreign gene or fragment of DNA (assumed as atypical in sequence characteristics) becomes compositionally more similar to the host genome with increasing residence time. This process has been called "amelioration," after reasoning that it makes a gene "better" (Lawrence and Ochman 1997), implying better translatability. This concept is a natural consequence of methods assuming that most foreign genes display mainly a poor-CU profile. However, as our results indicate, most foreign genes with poor CU are counterselected for successful integration, suggesting that CU amelioration might occur in a small fraction of genes. Of course, once integrated, foreign genes will be subjected to the same mutational biases as the rest of the indigenous genes, and if they had any atypical composition in terms other than CU, like GC content or oligonucleotide compositions (which are related to, but are not, CU), the mutational drift in the genome would "dilute" such differences without a dramatic impact on the CU of foreign genes relative to the genomic codon abundances. In fact, it has been observed that some genes may show rich CU but an average GC content (Garcia-Vallvé, Romeu, and Palau 2000; Garcia-Vallvé et al. 2003), and similarly we have observed genes that show deviant GC content but typical CU. Thus, the process is nothing else than the genetic drift that affects all genes within a genome, most probably making the genes fluctuate randomly within the typical CU area or safety zone without seriously affecting levels of expression.

A substantial number of prophage, transposase, and insertion sequence-related genes display low GC content and poor CU relative to chromosomal genes. However, this is not the case for most genes linked to mobile elements, as more than half of them show typical CRI in the host genome. For example, 60% of the genes in the plasmids of *A. tumefaciens* C58 and *M. loti* display typical to high CRI within their respective host genomes. Similarly, we identified 154 genes whose annotated function is directly related to mobile elements in *E. coli* K12 (e.g., phages, transposons, and plasmids) and found that 96 (62%) genes display typical CRI. This CU compatibility between genes that map to mobile elements and their host genomes is consistent with the previous observation that plasmids tend to show genome signatures similar to those of potential hosts (Campbell, Mrazek, and Karlin 1999) and with the hypothesis that these genome signatures might play an important role in HGT (Karlin 2001).

Concluding Remarks

We have shown that, despite the huge probability for a foreign gene to display poor CU, most detected HGTs involve genes with typical or rich CU. As there is no reason to assume that evolution today differs from what happened during most of the evolutionary history, successful HGTs should have involved ready-to-use genes. Although it is accepted that a substantial fraction of acquired genes might not be sufficiently atypical to be detected by most published methods (Lawrence and Ochman 2002), our results indicate that the great majority of recently acquired genes exhibit typical CU. It might be argued that foreign genes enter a genome with poor CU and then move to the typical CU level by means of "fast amelioration," thus escaping detection by our method. Although more detailed studies are required to fully address this issue, our results provide evidence against this alternative. For example, *T. maritima* still perceives most archaeal CXGs as typical genes. That is, the CRI of the archaeal genes calculated based on the codon abundances of *T. maritima* falls mainly within the typical-CRI (safety) zone, suggesting that if the genes could be presently transferred, they would arrive mainly with typical CU. This is also true for about 60% of all archaeal genes, not predicted as exchanged with *T. maritima*, in genomes such as *M. jannaschii*, *P. abyssi*, and *A. fulgidus* (see *Results*

above), which suggests that some potential archaeal donors currently show a strong CU compatibility with *T. maritima*, and any gene exchange with them would most likely involve typical-CRI genes.

The number of genes predicted here as involved in HGT events is quite small, but we are confident that they constitute a representative sample of true xenologs. Such reduced sample size is not unexpected given the stringent constraints imposed by our phylogenetic validations and given the fact that only those pairs of xenologs still presenting similar DNA sequence characteristics can clarify whether or not the CU of foreign genes is typical or not at the moment of introgression. Nonetheless, this methodology will benefit from the ever-growing number of sequenced genomes, as more pairs of xenologs will be detected. With a greater number of reliable xenologous genes we will be able to explore in greater detail the quantity and quality of lateral exchanges among genomes and thus to understand the behavior of HGT networks. The strategy presented herein provides a conceptual change in the way CU is used to gain a deeper knowledge of the processes involved in horizontal gene transfer.

Supplementary Material

Online Supplementary Material can be found at the journal's Web site (www.mbe.oupjournals.org).

Acknowledgments

We thank W. Lamboy, E. Morett, A. Garciarrubio, E. Merino, L. Martínez-Castilla, and two anonymous referees for valuable comments on the manuscript. A.M.-S. acknowledges a Ph.D. fellowship from CONACyT. This work has been supported by grant number 0028 from CONACyT to J.C.-V. We appreciate computer technical support from V. del Moral, E. Díaz, and C. Bonavides.

Literature Cited

Beaber, J. W., B. Hochhut, and M. K. Waldor. 2004. SOS response promotes horizontal dissemination of antibiotic resistance genes. Nature **427**:72–74.

Berg, O. G., and C. G. Kurland. 2002. Evolution of microbial genomes: sequence acquisition and loss. Mol. Biol. Evol. **19**:2265–2276.

Bernardo, J. O., and A. M. F. Smith. 1994. Bayesian theory. John Wiley and Sons, New York.

Brown, J. R. 2003. Ancient horizontal gene transfer. Nat. Rev. Genet. **4**:121–132.

Campbell, A., J. Mrazek, and S. Karlin. 1999. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proc. Natl. Acad. Sci. USA **96**:9184–9189.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. **17**:540–552.

Denamur, E., G. Lecointre, P. Darlu et al. (12 co-authors). 2000. Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. Cell **103**:711–721.

Ermolaeva, M. D., O. White, and S. L. Salzberg. 2001. Prediction of operons in microbial genomes. Nucleic Acids Res. **29**:1216–1221.

Felsenstein, J. 1989. PHYLIP: phylogeny inference package. Version 3.2. Cladistics **5**:164–166.

Garcia-Vallvé, S., E. Guzman, M. A. Montero, and A. Romeu. 2003. HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. Nucleic Acids Res. **31**:187–189.

Garcia-Vallvé, S., A. Romeu, and J. Palau. 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res. **10**:1719–1725.

Gogarten, J. P., W. F. Doolittle, and J. G. Lawrence. 2002. Prokaryotic evolution in light of gene transfer. Mol. Biol. Evol. **19**:2226–2238.

Gray, K. M., and J. R. Garey. 2001. The evolution of bacterial LuxI and LuxR quorum sensing regulators. Microbiology **147**:2379–2387.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52**:696–704.

Hanada, K., T. Ukita, Y. Kohno, K. Saito, J. Kato, and H. Ikeda. 1997. RecQ DNA helicase is a suppressor of illegitimate recombination in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **94**:3860–3865.

Hooper, S. D., and O. G. Berg. 2003. Duplication is more common among laterally transferred genes than among indigenous genes. Genome Biol. **4**:R48.

Ikemura, T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. **151**:389–409.

———. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. J. Mol. Biol. **158**:573–597.

Kaneko, T., Y. Nakamura, S. Sato et al. (24 co-authors). 2000. Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. DNA Res. **7**:331–338.

Karlin, S. 2001. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. Trends Microbiol. **9**:335–343.

Karlin, S., J. Mrázek, and A. M. Campbell. 1998. Codon usages in different gene classes of the *Escherichia coli* genome. Mol. Microbiol. **29**:1341–1355.

Koski, L. B., and G. B. Golding. 2001. The closest Blast hit is often not the nearest neighbor. J. Mol. Evol. **52**:540–542.

Koski, L. B., R. A. Morton, and G. B. Golding. 2001. Codon bias and base composition are poor indicators of horizontally transferred genes. Mol. Biol. Evol. **18**:404–412.

Kroll, J. S., K. E. Wilks, J. L. Farrant, and P. R. Langford. 1998. Natural genetic exchange between *Haemophilus* and *Neisseria*: intergeneric transfer of chromosomal genes between major human pathogens. Proc. Natl. Acad. Sci. USA **95**: 12381–12385.

Kyrpides, N. C., and G. J. Olsen. 1999. Archaeal and bacterial hyperthermophiles: horizontal gene exchange or common ancestry? Trends Genet. **15**:298–299.

Lawrence, J. G., and H. Ochman. 1997. Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. **44**: 383–397.

———. 1998. Molecular archaeology of the *Escherichia coli* genome. Proc. Natl. Acad. Sci. USA **95**:9413–9417.

———. 2002. Reconciling the many faces of lateral gene transfer. Trends Microbiol. **10**:1–4.

Majewski, J. 2001. Sexual isolation in bacteria. FEMS Microbiol. Lett. **199**:161–169.

Martin, W. 1999. Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. Bioessays **21**:99–104.

Médigue, C., T. Rouxel, P. Vigier, A. Henaut, and A. Danchin. 1991. Evidence for horizontal gene transfer in *Escherichia coli* speciation. J. Mol. Biol. **222**:851–856.

Moreno-Hagelsieb, G., and J. Collado-Vides. 2002*a*. A powerful non-homology method for the prediction of operons in prokaryotes. Bioinformatics **18**(Suppl 1):S329–S336.

———. 2002*b*. Operon conservation from the point of view of *Escherichia coli*, and inference of functional interdependence of gene products from genome context. In Silico Biol. **2**:87–95.

Moszer, I., E. P. Rocha, and A. Danchin. 1999. Codon usage and lateral gene transfer in *Bacillus subtilis*. Curr. Opin. Microbiol. **2**:524–528.

Needleman, S. B., and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48**:443–453.

Nelson, K. E., R. A. Clayton, S. R. Gill et al. (25 co-authors). 1999. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature **399**:323–329.

Nesbo, C. L., S. L'Haridon, K. O. Stetter, and W. F. Doolittle. 2001. Phylogenetic analyses of two "archaeal" genes in *Thermotoga maritima* reveal multiple transfers between archaea and bacteria. Mol. Biol. Evol. **18**:362–375.

Ochman, H., and U. Bergthorsson. 1998. Rates and patterns of chromosome evolution in enteric bacteria. Curr. Opin. Microbiol. **1**:580–583.

Olson, S. A. 2002. EMBOSS opens up sequence analysis. European molecular biology open software suite. Brief Bioinform. **3**:87–91.

Ragan, M. A. 2001*a*. Detection of lateral gene transfer among microbial genomes. Curr. Opin. Genet. Dev. **11**:620–626.

———. 2001*b*. On surrogate methods for detecting lateral gene transfer. FEMS Microbiol. Lett. **201**:187–191.

———. 2002. Reconciling the many faces of lateral gene transfer. Trends Microbiol. **10**:4.

Ramakrishnan, V., and S. W. White. 1998. Ribosomal protein structures: insights into the architecture, machinery and evolution of the ribosome. Trends Biochem. Sci. **23**:208–212.

Schaffer, A. A., L. Aravind, T. L. Madden, S. Shavirin, J. L. Spouge, Y. I. Wolf, E. V. Koonin, and S. F. Altschul. 2001. Improving the accuracy of PSI-Blast protein database searches with composition-based statistics and other refinements. Nucleic Acids Res. **29**:2994–3005.

Scott, K. P. 2002. The role of conjugative transposons in spreading antibiotic resistance between bacteria that inhabit the gastrointestinal tract. Cell Mol. Life Sci. **59**:2071–2082.

Sharp, P. M., and W. H. Li. 1987. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15**:1281–1295.

Smith, N. G., and A. Eyre-Walker. 2001. Why are translationally sub-optimal synonymous codons used in *Escherichia coli*? J. Mol. Evol. **53**:225–236.

Syvanen, M. 1994. Horizontal gene transfer: evidence and possible consequences. Annu. Rev. Genet. **28**:237–261.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25**:4876–4882.

Wang, B. 2001. Limitations of compositional approach to identifying horizontally transferred genes. J. Mol. Evol. **53**:244–250.

Wheeler, D. L., C. Chappey, A. E. Lash, D. D. Leipe, T. L. Madden, G. D. Schuler, T. A. Tatusova, and B. A. Rapp. 2000. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. **28**:10–14.

Wu, L., J. K. Karow, and I. D. Hickson. 1999. Genetic recombination: helicases and topoisomerases link up. Curr. Biol. **9**:R518–R520.

Zeigler, D. R. 2003. Gene sequences useful for predicting relatedness of whole genomes in bacteria. Int. J. Syst. Evol. Microbiol. **53**:1893–1900.

Peer Bork, Associate Editor

Accepted June 2, 2004

Anexo II

Nuestra actual era de ansiedad es, en gran parte, el resultado de intentar hacer el trabajo de hoy con las herramientas de ayer.

MARSHALL MCLUHAN

Se adjunta el artículo que fue publicado en la revista internacional *Journal of Statistical Software*, producto de un proyecto paralelo independiente a la tesis doctoral. Se reporta un método Bayesiano (al que se denominó BClass por sus siglas en ingles: *Bayesian Classifier*), basado en modelos mezcla, para relacionar (agrupar) entidades biológicas (e.g. genes) descritas por atributos cuya naturaleza matemática es muy heterogénea. Se utilizan varias distribuciones estadísticas (i.e. Normal, Poisson y Multinomial) para modelar los datos categóricos y continuos comúnmente empleados en las ciencias genómicas. El sistema calcula la probabilidad posterior de que cada entidad biológica pertenezca a un elemento (grupo) en el modelo mezcla. De esta manera, el conjunto original de variables biológicas heterogéneas es transformado en un conjunto de datos 100% homogéneo representado por las probabilidades posteriores de que cada entidad biológica pertenezca a cada uno de los grupos en la mezcla. Utilizado métodos MCMC estándar (Gibbs sampling y Metropolis–Hastings), se calcularon los momentos posteriores y probabilidades de agrupamiento. Dado que este método no requiere de la estimación de medidas de similitud o distancias, es especialmente adecuado para realizar minado de datos y descubrimiento de conocimiento den bases de datos biológicas.

Journal of Statistical Software

February 2005, Volume 13, Issue 2. *http://www.jstatsoft.org/*

BClass: A Bayesian Approach Based on Mixture Models for Clustering and Classification of Heterogeneous Biological Data

Arturo Medrano-Soto
Centro de Ciencias Genómicas

J. Andrés Christen
Centro de Investigación en Matemáticas

Julio Collado-Vides
Centro de Ciencias Genómicas

Abstract

Based on mixture models, we present a Bayesian method (called **BClass**) to classify biological entities (e.g. genes) when variables of quite heterogeneous nature are analyzed. Various statistical distributions are used to model the continuous/categorical data commonly produced by genetic experiments and large-scale genomic projects. We calculate the posterior probability of each entry to belong to each element (group) in the mixture. In this way, an original set of heterogeneous variables is transformed into a set of purely homogeneous characteristics represented by the probabilities of each entry to belong to the groups. The number of groups in the analysis is controlled dynamically by rendering the groups as 'alive' and 'dormant' depending upon the number of entities classified within them. Using standard Metropolis-Hastings and Gibbs sampling algorithms, we constructed a sampler to approximate posterior moments and grouping probabilities. Since this method does not require the definition of similarity measures, it is especially suitable for data mining and knowledge discovery in biological databases. We applied **BClass** to classify genes in **RegulonDB**, a database specialized in information about the transcriptional regulation of gene expression in the bacterium *Escherichia coli*. The classification obtained is consistent with current knowledge and allowed prediction of missing values for a number of genes.

BClass is object-oriented and fully programmed in Lisp-Stat. The output grouping probabilities are analyzed and interpreted using graphical (dynamically linked plots) and query-based approaches. We discuss the advantages of using Lisp-Stat as a programming language as well as the problems we faced when the data volume increased exponentially due to the ever-growing number of genomic projects.

Keywords: genetic databases, bioinformatics, MCMC, mixture models, clustering, data mining.

1. Introduction

The exponential growth of raw biological information represents an unprecedented challenge for biologists and bioinformaticians. Striking breakthroughs in biotechnology currently allow sequencing an average bacterial genome (the total DNA within an organism) in a matter of days. Since 1995 when the first complete sequence of a free-living bacterium *Haemophilus influenzae* was obtained (Flaischmann and *et al* 1995) to November 2004, there are already more than 200 complete genomes available, sampling the three domains of life, at The National Center for Biotechnology information (NCBI, `http://www.ncbi.nih.gov`), and more than 140 are still under way. Furthermore, high throughput experimental approaches have been developed to measure simultaneously the expression of all genes within an organism, both at the level of messenger RNA (Brown and Botstein 1999) and protein content (Anderson *et al.* 2000; Dutt and Lee 2000); such approaches have led to the emergence of the fields of transcriptomics and proteomics respectively. Statistical interpretation of transcriptome results is not a straightforward endeavor since different growth conditions, different RNA extraction procedures and different microarray-building systems hamper comparisons between experiments. Pitfalls, challenges and limits of these approaches have been adequately discussed elsewhere Danchin and Sekowska (2000). Experiments focused on individual biological systems, which involve a few genes, continue to appear in the literature and a substantial part of the resulting information is available in specialized databases. The inescapable consequence is that the pace at which raw experimental data is generated has greatly exceeded our ability to extract insightful knowledge from such information. For example, it is not trivial to infer from sets of coexpressed genes under a given experimental condition, which genes are corregulated, what are the regulatory proteins that regulate them, and even much more complicated, what is the network of regulatory interactions that produces the observed expression patterns. More robust methods and enhanced computational tools are required to come to grips with this problem by integrated analyses of heterogeneous data types.

We focused on the problem of clustering and knowledge discovery in **RegulonDB**, a biological database specialized in information about operon organization and transcriptional regulation of gene expression in the bacterium *Escherichia coli* (Salgado *et al.* 2004). The specific goal is to build a statistical framework that, hopefully, will allow uncovering previously unknown relationships among genes, given the diverse attributes that describe them. In our strategy we use a classical mixture model to tackle the problem of multivariate, heterogeneous classification and clustering. We call the resulting software **BClass**. Let $\boldsymbol{X} = (\boldsymbol{x}_{iv})$ be a (non-relational) database where $\boldsymbol{x}_{iv}$ represents attribute v of gene (entry) i; $i = 1, 2, \ldots, n$ and $v = 1, 2, \ldots, C$. Let also $\boldsymbol{x}_i$ be the ith row of $\boldsymbol{X}$; all the attributes for gene i. By heterogeneous databases we mean that the $\boldsymbol{x}_{iv}$'s may be very different in nature: continuous, discrete, categorical, etc. See Section 6 for an application example of **BClass** to **RegulonDB**.

A variety of clustering algorithms based on dissimilarity or distance measures are currently available. Methods for clustering and classification of heterogeneous, mixed, variables include converting variables to homogeneous types, analyzing variables separately and finding a weighted average of standardized dissimilarity measures (see, for example, Kaufman and Rousseau 1990; Gordon 1981). Regarding this latter method, many authors have suggested procedures to properly define the weights in a joint dissimilarity measure. However, as yet, there is the caveat that no standard method has been widely accepted and much heuristics is here needed (see Everitt 1993). As an alternative, mixture models have the advantage of

not relying either on distance measures nor on building (or attempting to build) a statistical model of the data analyzed (see Everitt 1993; McLachlan and Basford 1988) The now well known software **AutoClass** of Cheeseman and Stutz (1996) uses mixtures for clustering and classification in complex databases; it is largely based on the techniques presented by Titterington *et al.* (1985). **AutoClass** uses expected maximization (EM) approximations to find, mainly, maximum *a posteriori* (MAP) estimators (point estimators). In this paper, we build a mixture model for $\boldsymbol{X}$ using the assumption of conditional independence across elements in the mixture and attributes; this assumption has been shown both empirically and theoretically to be highly effective (Hand and Keming 2001). We embarked on doing a full Bayesian analysis using Markov Chain Monte Carlo (MCMC, see for example Gamerman 1997) methods to approximate complete posteriors, specifically, all grouping probabilities. This means that, in particular, given the number of groups or components in the mixture (J), we obtain a matrix $\boldsymbol{P} = (p_{ij})$ where p_{ij} is the posterior probability for entry i to belong to group j, $j = 1, 2, \ldots, J$. The whole process may be regarded as a transformation of $\boldsymbol{X}$ to $\boldsymbol{P}$, where now the "attributes" of each entry are its grouping probabilities p_{ij}'s. These new "attributes" are perfectly homogeneous. We then may use, as posterior interpretation tools, simple clustering techniques on $\boldsymbol{P}$ to find clusters. Therefore, both groups and posterior grouping probabilities (p_{ij}) are well defined mathematical objects, and "Clusters" are rather intuitive, and more interpretable, in terms of the database under consideration. Overall, we follow **AutoClass** philosophy of "automatic" classification, that is, setting the least number of parameters and relying on reasonable default values.

The problem of deciding the appropriate number of components in a mixture has been studied by many authors. However, there are only a handful of full Bayesian approaches that state a prior and calculate a posterior for the number of components J in the mixture. The difficulty here is that, as J varies, the dimension of the model changes, and standard MCMC approaches cannot handle chains of variable dimension. Philips and Smith (1996) propose using "jump diffusions", while Richardson and Green (1997) make an application in mixture models of the now popular "Reversible Jump" MCMC, developed by Green (1995). Stephens (1997) proposes yet another approach, based on point-process simulation, to deal with variable number of components. These strategies approach the problem in a full Bayesian setting being applications of general methods for variable dimension models and tend to become quite elaborate. Here, we tailored a procedure that allows handling different number of components, yet keeping J fixed. This is achieved using a vector of indicator ("dimensionality") variables that designate the "alive" and the "dormant" components. Using an alternative prior for the mixing probabilities that prefers parsimonious parameterizations, we construct a Markov Chain Monte Carlo (MCMC) relying on standard Gibbs sampling and Metropolis-Hastings algorithms (see Besag *et al.* 1995). It is worth noting that our method is suited for any set of component distributions, whereas the aforementioned approaches have been explored mainly for Normal mixtures.

The paper is organized as follows. In Section 2 we explain our model and the general form of the hierarchical structure. Specific priors for various distributions of gene attributes are explained in Section 3 and in Section 4 we discuss the posterior analysis of the MCMC output. In Section 5 we describe our general strategy to implement the system in Lisp-Stat. In Section 6 we apply **BClass** to the analysis of **RegulonDB** (Salgado *et al.* 2004) and some comparisons are made with the software **AutoClass**. Finally, a general discussion of the paper is given in Section 7.

2. The model

We use the standard mixture model with allocation variables as explained, for example, in Richardson and Green (1997). However, we add a vector of indicator variables to control the "alive" and the "dormant" groups in the mixture. That is, given a set of (unknown) parameters $\boldsymbol{h}$, the database $\boldsymbol{X}$ has distribution

$$f(\boldsymbol{X} \mid \boldsymbol{h}) = \sum_{j=1}^{J} \frac{\phi_j \pi_j}{\sum_{m=1}^{J} \phi_m \pi_m} f(\boldsymbol{X} \mid \boldsymbol{\theta}_j),$$

where $\pi_j \geq 0$, $\sum_{j=1}^{J} \pi_j = 1$, are the mixing probabilities, $f(\boldsymbol{X} \mid \boldsymbol{\theta}_j)$ are the within-group (component) distributions, and the indicator variables $\phi_j = 0, 1$ are the "dimensionality variables". Thus if $\phi_j = 0$, then group j is dormant (not considered), and if $\phi_j = 1$, group j is alive. We consider the number of components J to be fixed to a suitable large number and by integrating out the ϕ_j's we tackle the problem of "finding" the number of components in the mixture (the components that remained "alive"). We then have that $\boldsymbol{h} = (\boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\theta})$, where $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_J)$, $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_J)$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_J)$.

We assume that all the component distributions $f(\boldsymbol{x}_i \mid \boldsymbol{\theta}_j)$ belong to the same parametric family. We also assume conditional independence among components and attributes in the following sense,

$$f(\boldsymbol{X} \mid \boldsymbol{\theta}_j) = \prod_{i=1}^{n} f(\boldsymbol{x}_i \mid \boldsymbol{\theta}_j)$$

and

$$f(\boldsymbol{x}_i \mid \boldsymbol{\theta}_j) = \prod_{v=1}^{C} f(\boldsymbol{x}_{iv} \mid \boldsymbol{\theta}_{jv}),$$

where $\boldsymbol{\theta}_{jv}$ are the parameters needed for component j and attribute v. The above assumptions mean that we consider all attributes as conditionally independent *within each group*. This is a fairly reasonable assumption since, once all other sorts of variability are discarded, the internal group variability may be expected to behave as random, non-correlated, error. As supporting evidence, the studies of Hand and Keming (2001) presented an empirical and theoretical performance analysis of the assumption of conditional independence among attributes. They conclude that, although at first glance naive and most likely incorrect,

> The independence Bayes model seems often to perform surprisingly well.

(Hand and Keming 2001, p. 395). Part of the reason is that less parameters are needed to be estimated, thus outperforming models that consider interaction parameters, specially for several attributes. Furthermore, generally speaking, little is known about the databases studied and thus simple models need to be used (this approach is also taken in **AutoClass** and has given promising results, see Cheeseman and Stutz (1996). In order to consider non-independent attributes in the database, we could use multivariate distributions with some correlation structure. The generalization of the techniques developed here to use multivariate attributes is relatively straightforward (see Fraley and Raftery 1998, as an example using normal variables only). However, in this paper we take $f(\boldsymbol{x}_{iv} \mid \boldsymbol{\theta}_{jv})$ to be univariate, thus we write x_{iv}, a scalar. For clarification, say we have $C = 3$ attributes $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3})$ for each

gene i, and that these are Normal, Multinomial with four levels and Poisson. In such a case we would have $f(\boldsymbol{x}_i \mid \boldsymbol{\theta}_j) = f(x_{i1} \mid \mu_j, \sigma_j) f(x_{i2} \mid p_{j1}, p_{j2}, p_{j3}, p_{j4}) f(x_{i3} \mid \lambda_j)$ with the obvious notation.

Certainly, in general, we do not know which group each gene belongs to. Thus, for gene i, the group this gene belongs to is given by the latent allocation variable $J_i = 1, 2, \ldots, J$. Given these allocation variables, we have

$$f(\boldsymbol{x}_i \mid \boldsymbol{J}, \boldsymbol{\pi}, \boldsymbol{\theta}) = f(\boldsymbol{x}_i \mid \boldsymbol{\theta}_{J_i})$$

where $\boldsymbol{J}$ is the vector of J_i's. That is, given $\boldsymbol{J}$, $\boldsymbol{x}_i$ is drawn from its corresponding group J_i. For a given J our parameters are $(\boldsymbol{J}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\theta})$. We assume that $f(\boldsymbol{J}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\theta}) = f(\boldsymbol{J}, \boldsymbol{\pi}, \boldsymbol{\phi}) f(\boldsymbol{\theta})$ and *a priori*,

$$P(J_i = j \mid \boldsymbol{\pi}, \boldsymbol{\phi}) \propto \phi_j \pi_j$$

with $f(\boldsymbol{J} \mid \boldsymbol{\pi}, \boldsymbol{\phi}) = \prod_{i=1}^{n} f(J_i \mid \boldsymbol{\pi}, \boldsymbol{\phi})$, and we assume that $P[\phi_j = 1] = \alpha$ independently (α will be taken equal to $\frac{1}{2}$, non-informative). Therefore, to construct our prior we are using the decomposition

$$f(\boldsymbol{J}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\theta}) = f(\boldsymbol{J} \mid \boldsymbol{\pi}, \boldsymbol{\phi}) f(\boldsymbol{\pi}) f(\boldsymbol{\phi}) f(\boldsymbol{\theta}).$$

We take $f(\boldsymbol{\theta}) = \prod_{j=1}^{J} \prod_{v=1}^{C} f(\boldsymbol{\theta}_{jv})$ and the definition of $f(\boldsymbol{\theta}_{jv})$ is left for Section 3. Alternatively, $\boldsymbol{\phi}$ and $\boldsymbol{\pi}$ may be considered not independent *a priori* by, for example, taking $f(\boldsymbol{\phi}, \boldsymbol{\pi}) = f(\boldsymbol{\phi} \mid \boldsymbol{\pi}) f(\boldsymbol{\pi})$, and making ϕ_j to depend on π_j, somehow. However, this will only be relevant for the case when prior information is availbale to distinguish the π_j's, which is not the case for $f(\boldsymbol{\pi})$. An influence diagram (see, for example, Richardson and Green 1997) for our model is presented in Figure 1.

We leave the more mathematical aspects of defining $f(\boldsymbol{\pi})$ and the design of the Markov Chain Monte Carlo algorithm for the Appendix. We now turn to consider the distributions for the attributes in the data base.

3. Distributions for different attributes

In this section we present distributions for specific types of variables that are commonly used in biological databases. However, it should be clear from the type of hierarchical modeling used that any other type of distribution may also be considered, provided a default (proper) prior is stated and sampling from its unknown parameters is properly described. Proper priors are needed since we could encounter empty groups and in such a case we would need to sample from the prior itself, during the MCMC iterations. As far as the MCMC sampler is concerned, no further adjustments are required and the sampling scheme for the rest of the parameters remains the same. Moreover, the prediction section below regarding missing or unobserved values, is completely general and not solely restricted to the distributions presented herein.

3.1. Normal heterocedastic

With respect to Normal variables we have $\boldsymbol{\theta}_{jv} = (\mu_j, \lambda_j)$ and $x_{iv} \mid J_i = j, \boldsymbol{\mu}, \boldsymbol{\lambda} \sim N(\mu_j, \lambda_j)$ where μ_j is the mean and λ_j the precision (inverse of the variance). Let also $\boldsymbol{\mu}$ be the vector of μ_j's and $\boldsymbol{\lambda}$ the vector of λ_j's. Richardson and Green (1997) take $\mu_j \sim N(\mu_0, \lambda_0)$ and $\lambda_j \sim Ga(\alpha, \beta)$ independently for all j and μ_0, λ_0 and α fixed to some data dependent values.

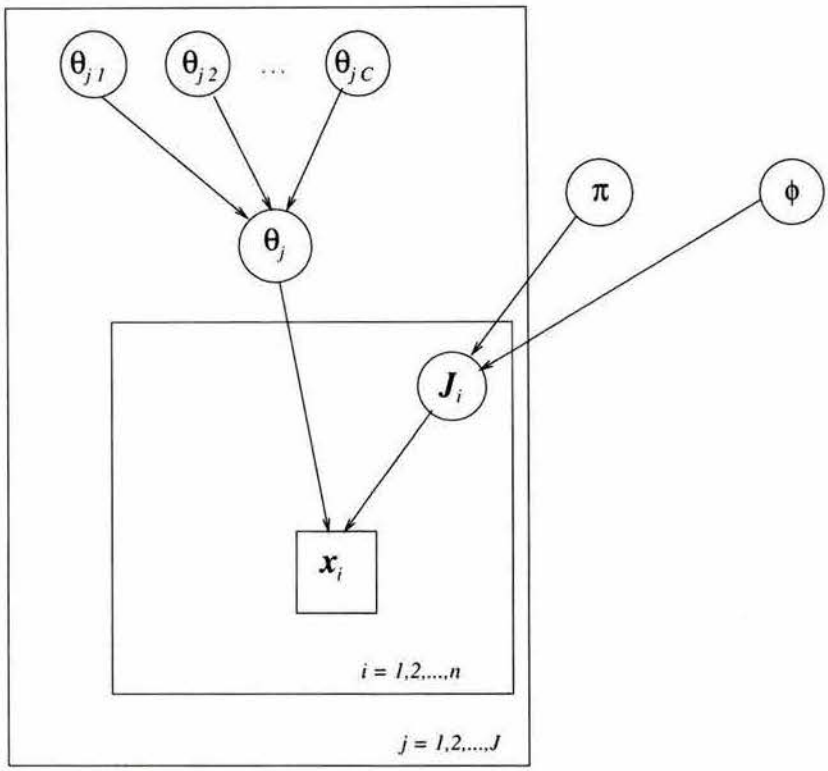

Figure 1: Directed acyclic graph for our model.

They consider a further hierarchy in the parameters by letting $\beta \sim Ga(g, h)$ with g and h fixed (Stephens 1997, also follows this approach), to construct a quasi-non-informative yet proper prior. We follow the same approach here. We fix the following parameters to $\mu_0 = \frac{a+b}{2}$, $g = 0.2$, $\alpha = 2$ $\lambda_0 = \frac{1}{(b-a)^2}$ and $h = \frac{100g}{\alpha(b-a)^2}$, where $a = \min x_{iv}$ and $b = \max x_{iv}$. Richardson and Green (1997) and Stephens (1997) point out that these values convey the belief that the "λ_j's are similar, without being informative about their absolute size". To update μ_j, λ_j and β, we simulate from their full conditionals (Gibbs kernel). That is, a $N(\mu_{n_j}, \lambda_{n_j})$, where $\lambda_{n_j} = \lambda_0 + n_j\lambda_j$ and $\mu_{n_j} = \lambda_{n_j}^{-1}(\lambda_0\mu_0 + n_j\lambda_{n_j}\sum_{J_i=j} x_{iv})$ for μ_j, a $Ga(\alpha + \frac{1}{2}n_j, \beta + \frac{1}{2}\sum_{J_i=j}(x_{iv} - \mu_j)^2)$ for λ_j and a $Ga(g + J\alpha, h + \sum_{j=1}^{J}\lambda_j)$ for β.

3.2. Normal homocedastic

We consider a homocedastic Normal variable, for which $x_{iv} \mid J_i = j, \boldsymbol{\mu}, \boldsymbol{\lambda} \sim N(\mu_j, \lambda)$, with the same precision λ across all groups. It is likely that in clustering problems we would prefer this variable, instead of a heterocedastic one (see Richardson and Green 1997) since it will tend to split the range of the data in more or less uniform intervals. Again, we take $\mu_j \sim N(\mu_0, \lambda_0)$ *a priori*, and $f(\lambda) \propto \lambda^{-1}$ as a prior for λ. To update μ_j we simulate from $N(\mu_{n_j}, \lambda_{n_j})$, where $\lambda_{n_j} = \lambda_0 + n_j\lambda$ and λ from a $Ga(\frac{1}{2}n, \frac{1}{2}\sum_{j=1}^{J}\sum_{J_i=j}(x_{iv} - \mu_j)^2)$ (the full conditionals), which are always proper although the prior for λ is not. As above, we take $\mu_0 = \frac{a+b}{2}$ (the data range mid point) and $\lambda_0 = \frac{1}{(b-a)^2}$, a large precision.

3.3. Multinomial

Regarding Multinomial attributes (eg. categorical), we assume that $x_{iv} = 1, 2, \ldots, L$ (that is, we translate labels into a numeric scale). The conjugate prior is a Dirichlet; however, in this case, there is not a unique standard reference prior, with $Di(1/L, \ldots, 1/L)$ being a common non-informative prior used. From simulated studies, where only one Multinomial variable is considered, we have seen that a $Di(1/2nL, \ldots, 1/2nL)$ leads to a mixture where only L components are effectively used and each component in fact takes only one level. We thus consider the latter as our default (sample size dependent) prior and simulate the vector of Multinomial probabilities in group j from its full conditional (Gibbs kernel), which is a $Di(1/2nL+n_{j1}, 1/2nL+n_{j2}, \ldots, 1/2nL+n_{jL})$, where $n_{jl} = |\{i : x_{iv} = l, J_i = j\}|$ (the current Multinomial counts for group j).

3.4. Poisson

A basic procedure to deal with some types of ordinal variables is to use a Poisson model. We assume that $x_{iv} = 0, 1, \ldots$. Given $x_{iv} \sim Po(\lambda_j)$ the standard reference prior is $f(\lambda_j) \propto \lambda_j^{-1/2}$, which is not proper. Following the hierarchical priors used for the Normal case, we take $\lambda_j \sim Ga(\alpha, \beta)$, fix $\alpha = 1$ and take $f(\beta) \propto \beta^{-1}$ (as a reference prior for β). We then simulate from the full conditional of λ_j which is a $Ga(\alpha + \sum_{J_i=j} x_{iv}, \beta + n_j)$ and we also simulate β from its full conditional (Gibbs kernels), that is a $Ga(J\alpha, \sum_{j=1}^{J} \lambda_j)$.

3.5. Prediction

In the context of MCMC, finding predictive distributions for missing (unobserved) attributes is straightforward. Simply, a missing attribute x_{iv} is taken as an unknown parameter and included in the Markov Chain simulation. From Section 2 we see that the full conditional of x_{iv} is $f(x_{iv} \mid \boldsymbol{\theta}_{J_i v})$, that is, sampling from the model used for attribute v. A "predictive" sample is then obtained for x_{iv}, which can be used either to approximate its predictive distribution or some point estimate like its predictive expectation.

4. Posterior analysis

As in any complex Bayesian analysis, an important problem is analyzing posterior information. **BClass** produces approximations for the posterior grouping probabilities $\boldsymbol{P} = (p_{ij}) = P(J_i = j \mid \boldsymbol{X})$, a $n \times J$ matrix, and posterior expectations for $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$, $E[\pi_j \mid \boldsymbol{X}]$ and $E[\boldsymbol{\theta}_{jv} \mid \boldsymbol{X}]$. As explained earlier one can regard the grouping probabilities as homogeneous attributes. For small n, it is possible to calculate a distance matrix (with simple Euclidean distance) and plot a dendrogram. However, even for moderate n (= 500), the dendrograms are difficult to read since most branches overlap. An alternative and simpler method is to plot the points $a_1 p_{i1} + a_2 p_{i2} + \cdots + a_J p_{iJ}$ where the a_j's are equally spaced along the unit circle. Thus the grouping probabilities for each entry i $(p_{i1}, p_{i2}, \ldots, p_{iJ})$ are mapped within the unit circle and similar grouping probabilities are plotted as nearby points (the contrary is not necessarily true though, and thus some care is needed in the interpretation of the resulting plots). These plots, hereafter referred to as "archipelago" plots, are used in the example to visually recognize clusters. We worked around the problem of cluster definition by plotting the grouping probabilities, say 10 times, with the order of the groups randomized. When

these plots are linked, clusters can be reliably defined as points located nearby in all the plots—false clusters always split apart when permuting the order of the groups.

With respect to the dimensionality variables, we only keep track of $k^{(l)} = \sum_{j=1}^{J} \phi_j^{(l)}$, the number of "alive" groups at pass l. Using the sample $k^{(l)}$ we approximate probabilities for the number of groups used. This is also done in the example.

5. BClass implementation in Lisp-Stat

BClass is fully implemented in Lisp-Stat for several reasons: (1) Lisp programming is remarkably straightforward; (2) it allows to focus on the problem at hand without much concern about the technical innards of the language; (3) most functions are vectorized; and (4) the Lisp-Stat's dynamic graphical-linking capabilities facilitate tremendously the exploratory analysis of both the input data and final results.

BClass was implemented following an object-oriented strategy. Every supported statistical model (i.e. Poisson, multinomial, normal homocedastic and normal heterocedastic) consists of a prototype that can be used to instantiate as many objects as necessary. Therefore, before running **BClass**, a preliminary analysis is required so as to decide which distribution type should be used to model each attribute.

The system includes a comprehensive tutorial plus a demo script to illustrate the readers how to load their own data into **BClass**, fine tune internal parameters (i.e. the number of groups J in the mixture, the number of iterations for the MCMC algorithm, etc.), how to start the classification process, plus how to save or reload the **BClass** output in order to postpone or continue a particular analysis. Two approaches (plot-based and query-based) are combined to define, interpret, and evaluate the quality of the resulting clusters, which are formed by observations sharing similar grouping probabilities. The system is open source code, it consists of 16 script files, the tutorial and the demo. **BClass** is available free of charge from the journal's web page and from the sites `http://www.cimat.mx/~jac/software` and `http://www.ccg.unam.mx/amedrano/BClass`.

6. Case study

In order to understand the instructive example here presented several, biological concepts are necessary. First, DNA is a linear sequence of four types of concatenated building blocks called nucleotides or bases: adenine (A), thymine (T), guanine (G) and cytosine (C). Structurally, DNA is a macromolecule with two complementary strands arranged in a double helix, each strand being read in opposite directions (forward or reverse) by the cellular machinery. Second, a gene is a DNA fragment that encodes the necessary information to produce proteins, which in turn are responsible for carrying out most of the cellular processes indispensable for sustaining life. Genes may be physically positioned in any of the two DNA strands and their length is the sum of As, Ts, Cs, and Gs in their sequence. Third, a bacterial chromosome is a circular self-replicating DNA sequence that contains in a linear array all or most of the genes. Fourth, transcriptional regulation refers to the molecular mechanism responsible for strategically expressing the genes required by the cell in order to survive environmental changes, reproduce, metabolize nutrients, etc. Fifth, the protein product of genes can be classified in several types, here we will deal only with four of them: enzyme (to accelerate biochemical

reactions), regulator (to control whether or not the protein product of other genes will be manufactured), transport (to take nutrients, toxins, etc in and out of the cell), and leader (specific sequences at the beginning of some genes that may function in targeting reactions or regulation). Sixth, gene function in this example refers to metabolic processes in which the genes participate (e.g. amino acid biosynthesis, energy production, cell division, etc.). These definitions, albeit sufficient, are by no means complete.

RegulonDB (Salgado *et al.* 2004) has information on transcriptional regulation for more than 2300 genes. We selected a set of 435 genes for which all attributes are completely described, with the exception of a few missing values (see below). We used the following attributes for the analysis: The DNA strand (forward or reverse), gene length in base pairs (bp), position within the chromosome (in minutes), gene type (enzyme, leader, regulator, transporter and miscellaneous), gene function (20 functional classes) and regulation mode (positive, negative or dual). These attributes, in turn, are modeled as Multinomial with 2 levels, Normal homocedastic, Normal homocedastic, Multinomial with 5 levels, Multinomial with 20 levels and Multinomial with 3 levels, respectively. There are 3 genes that have an unclassified or unknown function, and 11 genes with an unknown mode of regulation. Following the technique explained in Section 3.5 we predicted the values for these missing data. We considered a mixture of $(J =)30$ components.

Given the complexity of the multidimensional model entertained and the fact that Lisp-Stat is an interpreter, the convergence of the MCMC chain turned out to be slow. We took quite a long burn-in of 100,000 sweeps followed by a run of 50,000 sweeps. The π_j's were sorted and sampling was carried out every 5 sweeps. With the current **BClass** implementation in Lisp-Stat, these calculations took nearly 9 hours in a Intel Xeon processor (2.4 GHz) running Linux Red Hat 7.3. In this analysis we concentrated on calculating the grouping probabilities, that is, on obtaining the matrix $\boldsymbol{P} = (p_{ij})$. The archipelago plot for $\boldsymbol{P}$ (see Section 4) is shown in figure 2.

As explained above, using the sample $k^{(l)}$ we approximate probabilities for the number of groups used, see figure 3. We see that the most likely number of components is 29, and thus we have some confidence that we are using an appropriate number of groups.

So far, we have analyzed more than 15 clusters arising from the archipelago plots. For illustration, however, we present a brief analysis of the cluster shown in figure 2. This cluster has 13 genes. We see that all genes are regulated positively (their expression needs to be activated), all but two have a miscellaneous type, all are on the reverse strand, 12 genes participate in central intermediary metabolism (function 3) and the function of one gene is unknown. For this latter gene, *phnQ*, there is a predictive probability of more than 0.98 that its product is involved in central intermediary metabolism (function 3), as the rest of the genes in the cluster. Concerning their position within the chromosome the genes form two clusters: 12 genes form one cluster between minutes 92.93 and 93.11 and one gene, known as *gcvH*, is in minute 65.68. Regarding the size of the genes, they are in the range 309 to 1137 bp (small to average size), with the smallest gene being *gcvH*. In bacteria, a substantial fraction of genes is co-transcribed (expressed at the same time) defining the so-called "operons". All but *gcvH* belong to an operon containing 15 genes. This is the largest known operon in *E. coli* as most operons contain 2 or 3 genes. As far as we know, there is no reported evidence that these genes are functionally linked. Further theoretical and experimental research should be directed at studying the biological meaning of the association between the former gene and the latter operon. A preliminary transcriptome analysis of 4 growth conditions (i.e. heat shock,

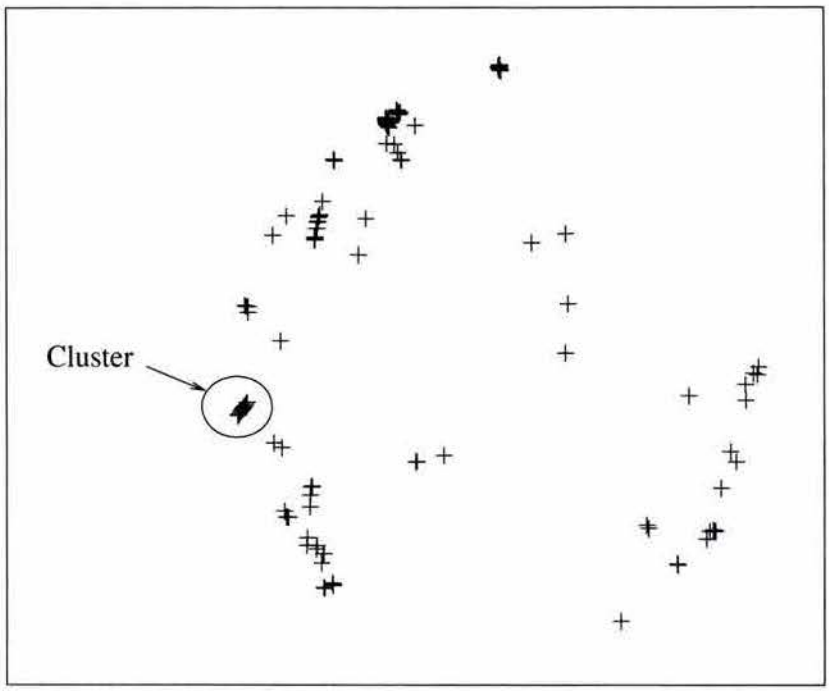

Figure 2: Archipelago plot for the 435 genes analyzed from the **RegulonDB** *E. coli* gene database.

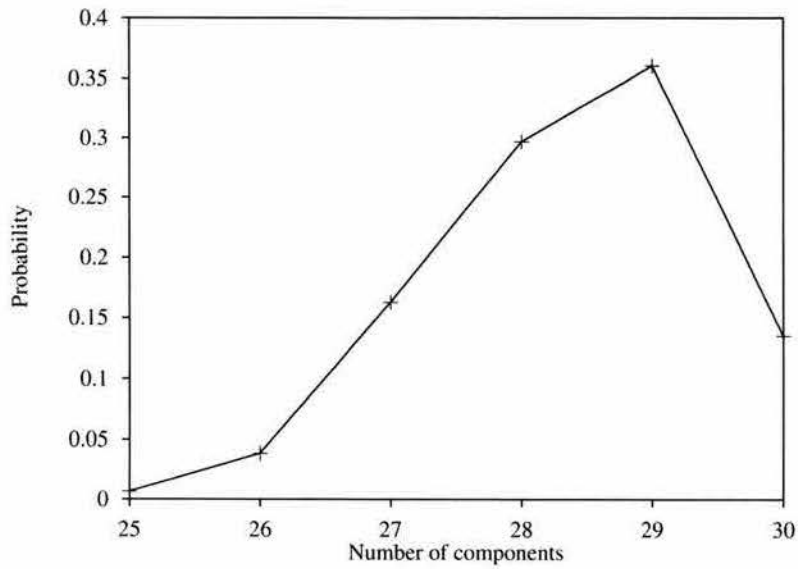

Figure 3: Posterior distribution for the number of "alive" components.

osmotic shock, minimal media, and IPTG) indicates that none of these genes change their expression significantly, so potential coexpression under other non-essayed growth conditions cannot be completely ruled out.

Besides the gene *phnQ* mentioned above, genes *yhdG* and *dsdX* have a predictive probability of nearly 1 that their product is involved in transport (function 14). Regarding the missing values for the type of regulation, gene *treB* has a probability of more than 0.99 of having a dual regulation type. The rest of the missing values analyzed do not have high predictive probabilities. It might be interesting to confirm these predictions with further experimental research.

For the sake of comparison, **AutoClass** was also applied to **RegulonDB**, for exactly the same genes and variables as above. Since **AutoClass** relies on MAP (point) estimates it only reports the most likely group each entry belongs to. **AutoClass** found 7 groups after 10,000 iterations (a minimum of 50 is recommended), however, none of these groups could be identified as one single cluster in the archipelago plot produced by **BClass**. Moreover, the groups produced by **AutoClass** were split into several well compacted more interpretable clusters by **BClass**. Indeed, it is not surprising that we can extract more significant information out of **BClass** since we do have access to the whole vector of grouping probabilities for each entry and not just the most likely group. More comprehensive evaluations are needed to fully address the issue of comparing the performance and capabilities of each software.

7. Discussion

Here we present a general tool for Bayesian classification of genetic databases using mixture models. The methodology is "open" in the sense that, in principle, any set or combination of quantitative genetic attributes may be analyzed, with few new technicalities to be taken care of.

As explained in Section A.2, without any identifiability constraint, the posterior distribution has $J!$ symmetric components. If a MCMC sampler is run without such constraints, samples will be drawn from any of these components. The output of such sampler should be analyzed with much care, avoiding mixing samples from different components. By post-processing the MCMC output, Stephens (1997) proposes two methods to "relabel" a sample from a mixture model. Stephens' second method could be applied here since it is independent of the type of component distributions used. However, this method post-processes the probabilities of the full conditionals for each J_i in every MCMC iteration. This involves keeping s matrices of size $n \times J$, where s is the MCMC sample size; a formidable task even for moderate n. The ordering constraint in the π_j's followed in this paper does avoid some of the label switching problems, provided that the groups do not have similar π_j's. The label switching problem has just recently been addressed, and thus further research efforts are required to find a simpler on-line solution.

The number of components used in the mixture is controlled by the indicator dimensionality variables ϕ_j's, while the entropy prior on the mixing probabilities enable the method to obtain parsimonious parametrizations. We use standard Metropolis-Hastings, which leads to reasonable simple algorithms, in contrast with the complexity of methods proposed elsewhere (Philips and Smith 1996; Richardson and Green 1997; Stephens 1997). Moreover, our approach is independent of the families of component distributions used, making unnecessary

any fine tuning of details for each new distribution introduced.

As discussed above, the rationale behind **BClass** offers several advantages, particularly in the case of genetic databases, compared to other clustering techniques. These ideas should be useful and robust for applications in diverse areas of research such as satellite imaging, social sciences, medicine etc. besides biology.

As far as we know, mixture models have been applied to biology at the level of morphological traits (na and noz 1998) but not to mine databases in molecular biology. Bayesian statistics has been certainly used in Bioinformatics in learning strategies to identify common motifs in related sequences (Neuwald *et al.* 1995), in sequence alignment (Zhu *et al.* 1998), phylogenetics (Lewis and Swofford 2001), as well as in transcriptome analysis (Baldi and Long 2001). Given the explosion and growth of biological databases (see the January issues of the journal Nucleic Acids Research where biological databases are presented), it is reasonable to foresee the increasing importance of mixture models as a tool for clustering analysis and biological knowledge discovery. A well known example of the contribution of clustering techniques to molecular biology is the determination of 3 general codon usage classes in *E. coli* (Médigue *et al.* 1991). These major codon usage groups have been observed in other bacteria as well, which fueled the development of improved computational methods for gene prediction (Borodovsky *et al.* 1995). More recently, the emergence of post-genomic experimental tools has generated an outburst of large data sets of expression profiles for all genes within complete genomes. The virtues of clustering analysis in functional genomics have been clearly illustrated by Eisen *et al.* (1998). Currently, clustering techniques are routinely used in transcriptome analysis to discover genes that might contribute to disease, identify potential drug targets, and sets of corregulated genes that will play an essential role in the eventual characterization of complete genomic regulatory networks (see D'Haeseleer *et al.* 2000). Even though expression values are homogeneous, the methodology here presented permits an integrated analysis through the inclusion of other additional gene attributes. For instance, **BClass** is able to analyze simultaneously transcriptome data, functional categories, regulation modes, position within the chromosome and/or other biological attributes that could strategically improve the search for sets of co-regulated genes. Future work with **BClass** shall illustrate the extent to which heterogeneous classification impinges upon integrative genomic analyses.

The application of **BClass** to a data subset of **RegulonDB** is presented here as a preliminary example. However, it was apparent that the current list-based implementation of **BClass** in Lisp-Stat, albeit functional, is not adequate to analyze large data sets because the code is interpreted. Given that the time required for the execution of the MCMC algorithm is a function of the number of parameters and the number of observations, we have re-implemented the whole system using an array-based approach and a commercial version of common Lisp (Allegro), in order to compile the source code and obtain a fast binary stand-alone application. Here we concentrated on efficiency issues to minimize time-consuming bottlenecks in the code. However, although the overall **BClass** speed was significantly increased, it is not fast enough to handle thousands of entries in a few minutes, often requiring several hours or even days. Given this situation, we are currently working on a parallel version of **BClass** that will be able to run much faster.

Acknowledgements

We thank an anonymous referee for valuable criticisms to the manuscript. AM-S acknowledges a Ph.D fellowship from CONACyT. This work has been supported by grant number 0028 from CONACyT to JC-V. We appreciate computer technical support by V. del Moral and C. Bonavides.

References

Anderson N, Matheson A, Steiner S (2000). "Proteomics: Applications in Basic and Applied Biology." *Current Opinion Biotechnology*, **11**, 408–412.

Baldi P, Long A (2001). "A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized t-Test and Statistical Inferences of Gene Changes." *Bioinformatics*, **17**, 509–519.

Bernardo J, Smith A (1994). *Bayesian Theory*. John Wiley & Sons, Chichester.

Besag J, Green P, Higdon D, Mengersen K (1995). "Bayesian Computation and Stochastic Systems." *Statistical Science*, **10**(1), 3–66.

Borodovsky M, McIninch J, Koonin E, Rudd K, Médigue C, Danchin A (1995). "Detection of New Genes in a Bacterial Genome using Markov Models for three Gene Classes." *Nucleic Acids Research*, **23**(17), 3554–3562.

Brown P, Botstein D (1999). "Exploring the New World of the Genome with DNA Microarrays." *Nature Genetics*, **21**, 33–37.

Cheeseman P, Stutz J (1996). "Bayesian Classification (**AutoClass**): Theory and Results." In PS U Fayyad G Piatetsky-Shapiro, R Uthurusamy (eds.), "Advances in Knowledge Discovery and Data Mining," pp. 153–180. AAAI Press/MIT Press.

Danchin A, Sekowska A (2000). "Expression Profiling in Reference Bacteria: Dreams and Reality." *Genome Biology*, **1**(4), 1024.

D'Haeseleer P, Liang S, Somogyi R (2000). "Genetic Network Inference: From Co-expression Clustering to Reverse Engineering." *Bioinformatics*, **16**, 707–726.

Dutt M, Lee K (2000). "Proteomic Analysis." *Current Opinion Biotechnology*, **11**, 176–179.

Eisen M, Spellman P, Brown P, Botstein D (1998). "Cluster Analysis and Display of Genome-wide Expression Patterns." *Proceedings of the National Academy of Sciences, USA*, **95**(25), 14863–14868.

Everitt B (1993). *Cluster Analysis*. Arnold, London.

Flaischmann R, *et al* (1995). "Whole Genome Random Sequencing and Assembly of *Haemophilus Influenzae*." *Science*, **269**, 496–512.

Fraley C, Raftery A (1998). "How Many Clusters? Which Clustering Method? Answer via Model-based Cluster Analysis." *The Computer Journal*, **41**(8), 579–587.

Gamerman D (1997). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference.* Chapman & Hall, London.

Gordon A (1981). *Classification.* Chapman & Hall, London.

Green P (1995). "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination." *Biometrika*, **82**, 711–732.

Hand Y, Keming Y (2001). "Idiot's Bayes—Not so Stupid after all?" *International Statistical Review*, **69**, 385–398.

Kaufman L, Rousseau P (1990). *Finding Groups in Data.* John Wiley & Sons, New York.

Lewis P, Swofford D (2001). "Back to the Future: Bayesian Inference Arrives in Phylogenetics." *Trends in Ecology & Evolution*, **16**(11), 600–601.

McLachlan G, Basford K (1988). *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, New York.

Médigue C, Rouxel T, Vigier P, Henaut A, Danchin A (1991). "Evidence for Horizontal Gene Transfer in *Escherichia Coli* Speciation." *Journal of Molecular Biology*, **222**(4), 851–856.

na JA, noz MM (1998). "Mixture Analysis in Biology: Scope and Limits." *Journal of Theoretical Biology*, **191**, 341–344.

Neuwald A, Liu J, Lawrence C (1995). "Gibbs Motif Sampling: Detection of Bacterial Outer Membrane Protein Repeats." *Protein Science*, **4**(8), 1618–1632.

Philips D, Smith A (1996). "Bayesian Model Comparison via Jump Diffusions." In SR W R Gilks, D Spiegelhalter (eds.), "Practical Markov Chain Monte Carlo," pp. 441–446. Chapman & Hall, London.

Richardson S, Green P (1997). "On Bayesian Analysis of Mixtures with an Unknown Number of Components." *Journal of the Royal Statistal Society, Series B*, **59**(4), 731–792.

Roeder K, Wasserman L (1997). "Practical Bayesian Density Estimation Using Mixtures of Normals." *Journal of the American Statistal Association*, **92**(439), 894–902.

Salgado H, Gama-Castro S, Martínez-Antonio A, Díaz-Peredo E, Sánchez-Solano F, Peralta-Gil M, García-Alonso D, Jiménez-Jacinto V, Santos-Zavaleta A, Bonavides-Martínez C, Collado-Vides J (2004). "**RegulonDB** (version 4.0): Transcriptional Regulation, Operon Organization and Growth Conditions in *Escherichia Coli* K-12." *Nucleic Acids Research*, **32**, D303–D306.

Stephens M (1997). *Bayesian Methods for Mixtures of Normal Distributions.* Ph.D. thesis, Magdalen College, Oxford.

Titterington D, Smith A, Makov U (1985). *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons, New York.

Zhu J, Liu J, Lawrence C (1998). "Bayesian Adaptive Sequence Alignment Algorithms." *Bioinformatics*, **14**(1), 25–39.

A. Technical considerations

A.1. Identifiability and the entropy prior

The problem of identifiability has been discussed extensively in the literature of mixture models (see, for example, Titterington *et al.* 1985). The problem may be described in the following way. Given two sets of parameters $\boldsymbol{h_1}$ and $\boldsymbol{h_2}$ for our mixture model $f(\boldsymbol{X} \mid \boldsymbol{h})$, when is it the case that $f(\boldsymbol{X} \mid \boldsymbol{h_1}) = f(\boldsymbol{X} \mid \boldsymbol{h_2}) \Leftrightarrow \boldsymbol{h_1} = \boldsymbol{h_2}$? (in such circumstances $\boldsymbol{h_1}$ uniquely describes the model and there are no alternative parametrizations). Obviously, if we have $\boldsymbol{h}$ and permute the indices j in $\boldsymbol{h}$, say $\sigma(\boldsymbol{h})$, we have $f(\boldsymbol{X} \mid \boldsymbol{h}) = f(\boldsymbol{X} \mid \sigma(\boldsymbol{h}))$. To avoid this, we only consider labelings in which $\pi_1 \leq \pi_2 \leq \ldots \leq \pi_J$. Now, it is well known, for example, that a mixture of Normals is identifiable, that is $f(\boldsymbol{X} \mid \boldsymbol{h_1}) = f(\boldsymbol{X} \mid \boldsymbol{h_2}) \Leftrightarrow \boldsymbol{h_1} = \boldsymbol{h_2}$ (besides, indeed, label permutations). On the contrary, it is easy to see that in general a mixture of Multinomials is not identifiable (see Titterington *et al.* 1985, p. 35) in which case we have a set H such that $f(\boldsymbol{X} \mid \boldsymbol{h_1}) = f(\boldsymbol{X} \mid \boldsymbol{h_2})$ for $\boldsymbol{h_1}, \ \boldsymbol{h_2} \in H$. When analyzing heterogeneous databases we will be dealing with Multinomial component distributions and, in general, we do not know which combinations of component distributions or even which families of distributions will be used in the analysis of specific databases. Moreover, even if we were dealing with an identifiable mixture we may still have the problem of *sub-identifiability*. For example, a mixture with five components, $J = 5$, but only four effective groups, $\phi_5 = 0$, could just as well be described with $\phi_5 = 1$, 5 groups. This illustrates two nested models that describe the data equally well.

Therefore, non-identifiability or sub-identifiability (or both) results in a likelihood $f(\boldsymbol{X} \mid \boldsymbol{h})$ with large flat areas. Thus, choosing among alternative parametrizations, could and should be done with an appropriate prior. Suppose then that $\boldsymbol{h_1}$ and $\boldsymbol{h_2}$ are two alternative parametrizations, how would we choose *a priori* among them? Indeed, we would like parsimonious parametrizations, using the least number of groups, or more "compact" groups. In mathematical terms we say that we select the parametrization that has the larger information content (less entropy) in its mixing probabilities. Thus if $\sum_{j=1}^{J} \pi_j^1 \log \pi_j^1 > \sum_{j=1}^{J} \pi_j^2 \log \pi_j^2$ we prefer $\boldsymbol{h_1}$ from $\boldsymbol{h_2}$ (using superindices to distinguish the π_j's).

Finally, to express the above in terms of a prior for $\boldsymbol{\pi}$ we propose the *entropy* prior

$$f(\boldsymbol{\pi}) \propto \exp\left\{\sum_{j=1}^{J} \pi_j \log \pi_j\right\}, \tag{1}$$

for $0 < \pi_1 \leq \pi_2 \leq \ldots \leq \pi_J < 1$ and $\sum_{j=1}^{J} \pi_j = 1$. Thus parametrizations with a higher information content in the mixing probabilities π_j's will have higher prior probability density. From the properties of the information score (see Bernardo and Smith 1994, p. 79) note that the kernel of $f(\boldsymbol{\pi})$ is bounded by 1, therefore $f(\boldsymbol{\pi})$ is well defined.

The natural conjugate prior for $\boldsymbol{\pi}$ is a Dirichlet, the prior used for $\boldsymbol{\pi}$ in virtually all Bayesian mixture model analysis. The entropy prior in (1) is not conjugate but still has a convenient form regarding the MCMC calculations that follow (We may think of a more general form for (1), say $f(\boldsymbol{\pi}) \propto \exp\{\sum_{j=1}^{J}(\alpha_j - 1)\log \pi_j + \gamma_j \pi_j \log \pi_j\}$. This distribution has as a special case the Dirichlet and would be a conjugate prior for $\boldsymbol{\pi}$. However, we decided to use (1) since it represents clearly the information we are trying to convey.).

A.2. MCMC

It is routine in almost any modern Bayesian analysis to use Markov Chain Monte Carlo (MCMC) methods to approximate posterior densities; for a review of the subject see Besag *et al.* (1995). We design our MCMC sampler in the following way. We simulate in turn the parameters $\boldsymbol{\theta}_{jv}$, for $j = 1, 2, \ldots, J$, $v = 1, 2, \ldots, C$ from their full conditionals. The allocation variables J_i's are simulated using a Metropolis step and we simulate $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$ in a single stage using a Metropolis-Hastings step.

It is easy to see that the full conditionals for $\boldsymbol{\theta}_{jv}$ are proportional to $\prod_{J_i=j} f(\boldsymbol{x}_i \mid \boldsymbol{\theta}_{jv}) f(\boldsymbol{\theta}_{jv})$. This means that $\boldsymbol{\theta}_{jv}$ will be drawn as if from a posterior distribution using the likelihood $\prod_{J_i=j} f(\boldsymbol{x}_i \mid \boldsymbol{\theta}_{jv})$ and prior $f(\boldsymbol{\theta}_{jv})$. We will concentrate on conjugate forms for $f(\boldsymbol{\theta}_{jv})$ and, in general, $\boldsymbol{\theta}_{jv}$ will be easy to simulate. There is, however, the difficulty that improper priors may not be used (see Roeder and Wasserman 1997). The definition of $f(\boldsymbol{\theta}_{jv})$ for specific type of variables is left for Section 3.

With respect to the allocation variables $\boldsymbol{J}$, we make a proposal J_i' for J_i with $P(J_i' = j) \propto \phi_j$ (that is, uniformly from the alive components). This represents a Metropolis (symmetrical) proposal and it is not difficult to see that its acceptance probability is

$$\min\left(1, \frac{f(\boldsymbol{x}_i \mid \boldsymbol{\theta}_{J_i'}) \pi_{J_i'}}{f(\boldsymbol{x}_i \mid \boldsymbol{\theta}_{J_i}) \pi_{J_i}}\right).$$

In our experience, both with simulated and real data, we have seen that this Metropolis step has better mixing than simulating directly from the full conditional of J_i, with the added benefit that only two evaluations of the likelihood are involved.

We construct proposals for $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$ to be accepted or rejected through a Metropolis-Hastings acceptance probability. Let $n_j = |\{i : J_i = j\}|$. We use a Dirichlet $Di(n_1+1, n_2+1, \ldots, n_J+1)$ to simulate a proposal $\boldsymbol{\pi}' = (\pi_1', \pi_2', \ldots, \pi_J')$ for $\boldsymbol{\pi}$. At a first step we ignore the ordering imposed on the π_j's. Due to alternative relabelings, the posterior distribution will have $J!$ symmetric components. Given a simulated value $\boldsymbol{h}^*$ from the posterior, a relabeling of $\boldsymbol{h}^*$ may be considered to be a simulated value from the posterior. We therefore run the MCMC sampler and after a burn-in, when samples may be viewed as drawn from the posterior, we relabel the components to have the correct ordering in the π_j's. After the burn-in we relabel the samples every 5 or 10 passes and use only those samples both to ensure the correct ordering in the π_j's and avoid correlation. Thus in what follows we take the prior for $\boldsymbol{\pi}$ as in (1) ignoring the ordering constraint.

We simulate a proposal $\boldsymbol{\phi}' = (\phi_1', \phi_2', \ldots, \phi_J')$ for $\boldsymbol{\phi}$, independently of $\boldsymbol{\pi}$, using $P(\phi_j' = 1 \mid n_j > 0) = 1$ and $P(\phi_j' = 1 \mid n_j = 0) = \beta$, for some suitable proposal probability β. Note that we are only considering the "birth" or "death" of a group $\phi_j' = 1, 0$ when the group is empty, $n_j = 0$, since a proposal $\phi_j' = 0$ given $n_j > 0$ has zero acceptance probability.

It is proved below that the acceptance ratio for this proposal is

$$A = \exp\left\{ u \log\frac{\alpha}{1-\alpha} + w \log\frac{\beta}{1-\beta} + \sum_{j=1}^{J} \pi_j' \log \pi_j' - \pi_j \log \pi_j \right\} \tag{2}$$

where $u = \sum_{j=1}^{J} \phi_j' - \phi_j$ and $w = \nu_0 - \nu_0'$, where $\nu_0 = |\{j : n_j = 0, \phi_j = 0\}|$ and $\nu_0' = |\{j : n_j = 0, \phi_j' = 0\}|$. $\boldsymbol{\pi}$' and $\boldsymbol{\phi}$' are then accepted with probability $\min(1, A)$. We see that it

is not difficult to simulate the proposals and that the acceptance ratio has a rather compact form and is quite simple to calculate.

A.3. Proof of (2)

As explained in Section 2 we use a Dirichlet $Di(n_1+1, n_2+1, \ldots, n_J+1)$ to simulate a proposal $\boldsymbol{\pi}'$ for $\boldsymbol{\pi}$ and we simulate a proposal $\boldsymbol{\phi}'$ for $\boldsymbol{\phi}$, independently of $\boldsymbol{\pi}$, using $P(\phi'_j = 1 \mid n_j > 0) = 1$ and $P(\phi'_j = 1 \mid n_j = 0) = \beta$. For the transition kernel we have that $K\{(\boldsymbol{\pi}, \boldsymbol{\phi}), (\boldsymbol{\pi}', \boldsymbol{\phi}')\} = k_1(\boldsymbol{\pi}')k_2(\boldsymbol{\phi}')$ and $k_1(\boldsymbol{\pi}') \propto \exp\left(\sum_{j=1}^{J} n_j \log \pi'_j\right)$ and $k_2(\boldsymbol{\phi}') = \exp\left(\nu'_1 \log\beta + \nu'_0 \log\beta\right)$, where $\nu'_h = |\{j : n_j = 0, \phi'_j = h\}|$, $h = 0, 1$. Noting that $f(\boldsymbol{\pi}, \boldsymbol{\phi} \mid \boldsymbol{X}, \boldsymbol{J}, \boldsymbol{\theta}) \propto f(\boldsymbol{J} \mid \boldsymbol{\pi}, \boldsymbol{\phi}) f(\boldsymbol{\pi}) f(\boldsymbol{\phi})$ and using that $f(\boldsymbol{J} \mid \boldsymbol{\pi}, \boldsymbol{\phi}) = \exp\left(\sum_{j=1}^{J} n_j \log \pi_j\right)$ for $\phi_j = 1$ such that $n_j > 0$ and zero otherwise, we see that the likelihood is canceled out with $k_1(\boldsymbol{\pi})$ and what is left is the prior ratio and the ratio for $k_2(\boldsymbol{\phi})$. Therefore we obtain

$$
\begin{aligned}
A \;=\; & \exp\left(\sum_{j=1}^{J} \pi'_j \log \pi'_j - \pi_j \log \pi_j\right) \\
& \exp\left(\sum_{j=1}^{J} (\phi'_j - \phi_j)\log\alpha + (\phi_j - \phi'_j)\log(1-\alpha)\right) \\
& \exp\left((\nu_1 - \nu'_1)\log\beta + (\nu_0 - \nu'_0)\log(1-\beta)\right),
\end{aligned}
$$

with the equivalent definition for ν_1 and ν_0. Letting $n_0 = |\{j : n_j = 0\}|$ we have that $\nu_1 = n_0 - \nu_0$ and it is easy to see that $\nu_0 - \nu'_0 = \nu'_1 - \nu_1 = v$; (2) follows immediately.

Affiliation:

Arturo Medrano-Soto
Program of Computational Genomics
Centro de Ciencias Genómicas (UNAM)
Ave. Universidad s/n, Col. Chamilpa, A.P. 565-A
62100 Cuernavaca, Morelos, Mexico.
E-mail: amedrano@ccg.unam.mx

J. Andrés Christen
Program of Probability and Statistics
Centro de Investigación en Matemáticas, A.C.
A.P. 402,
36000 Guanajuato, Gto., Mexico.
E-mail: jac@cimat.mx

Julio Collado-Vides
Program of Computational Genomics
Centro de Ciencias Genómicas (UNAM)
E-mail: collado@ccg.unam.mx

Journal of Statistical Software
February 2005, Volume 13, Issue 2.
http://www.jstatsoft.org/

Submitted: 2004-04-29
Accepted: 2004-12-20

Bibliografía

1. Medrano-Soto, A, Moreno-Hagelsieb, G, Vinuesa, P, et al. (2004). Successful lateral transfer requires codon usage compatibility between foreign genes and recipient genomes. Mol Biol Evol **21**(10):1884-94.
2. Medrano-Soto, A, Christen, JA y Collado-Vides, J. (2005). BClass: A Bayesian approach based on mixture models for clustering and classification of heterogeneous biological data. Journal of Statistical Software **13**(2):1-18.
3. Fitch, WM. (2000). Homology a personal view on some of the problems. Trends in Genetics **16**(5):227-31.
4. Mereschkowsky, C. (1905). Uber Natur und Ursprung der Chromatophoren in Pflanzenteilen. Biol. Zentrabl. **25**:593-635.
5. Wallin, JC (1928) *Symbioticism and the origin of the species.* (Baillere, Tindall & Cox, London).
6. Margulis, L. (1971). Symbiosis and evolution. Sci Am **225**(2):48-57.
7. Woese, CR. (1977). Endosymbionts and mitochondrial origins. J Mol Evol **10**(2):93-6.
8. Stroun, M, Anker, P y Auderset, G. (1970). Natural release of nucleic acids from bacteria into plant cells. Nature **227**(5258):607-8.
9. Nester, EW y Kosuge, T. (1981). Plasmids specifying plant hyperplasias. Annu Rev Microbiol **35**:531-65.
10. Zhu, J, Oger, PM, Schrammeijer, B, et al. (2000). The bases of crown gall tumorigenesis. J Bacteriol **182**(14):3885-95.
11. Ochia, K, Yamanaka, T, Kimura, K, et al. (1959). Inheritance of drug resistance (and its transfer) between Shigella strains and between Shigella and *E. coli* Strains. Nihon Iji Shimpo **1861**:34.
12. Akiba, T, Koyama, K, Ishiki, Y, et al. (1960). The mechanism of the development of multiple drug-resistant clones of Shigella. Jpn J. Microbiol. **4**:219.
13. Went, FW. (1971). Parallel evolution. Taxon **20**:197-226.
14. Krassilov, VA. (1977). The origin of angiosperms. Bot. Rev. **43**:143-176.
15. Anderson, NG. (1970). Evolurionary significance of virus infection. Nature **227**:1346-1347.
16. Reanney, D. (1976). Extrachromosomal elements as possible agents of adaptation and development. Bacteriol. Rev. **40**:552-590.
17. Hartman, H. (1976). Speculation on viruses, cells and evolution. Evolution Theory **3**:159-163.
18. Struhl, K, Cameron, JR y Davis, RW. (1976). Functional genetic expression of eukaryotic DNA in *Escherichia coli*. Proc. Natl. Acad. Sci, **73**(5):1471-1475.
19. Davies, J y Jiménez, A. (1980). A new selective agent for eukaryotic cloning vectors. Am J Trop Med Hyg **29**(5 Suppl):1089-92.
20. Palmiter, RD, Norstedt, G y Gelinas, RE. (1983). Metallothionein-human GH fussion genes stimulate growth of mice. Science **222**(4625):809-814.
21. Syvanen, M. (1985). Cross-species gene transfer; implications for a new theory of evolution. J Theor Biol **112**(2):333-43.
22. Grantham, R, Gautier, C, Gouy, M, et al. (1980). Codon catalog usage and the genome hypothesis. Nucleic Acids Res **8**(1):r49-r62.
23. Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol **151**(3):389-409.
24. Ikemura, T. (1982). Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. J Mol Biol **158**(4):573-97.
25. Robinson, M, Lilley, R, Little, S, et al. (1984). Codon usage can affect efficiency of translation of genes in *Escherichia coli*. Nucleic Acids Res **12**(17):6663-71.
26. Chen, GT y Inouye, M. (1994). Role of the AGA/AGG codons, the rarest codons in global gene expression in *Escherichia coli*. Genes Dev **8**(21):2641-52.
27. Kane, JF. (1995). Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. Curr Opin Biotechnol **6**(5):494-500.
28. Felmlee, T, Pellett, S y Welch, RA. (1985). Nucleotide sequence of an *Escherichia coli* chromosomal hemolysin. J Bacteriol **163**(1):94-105.

29. Sharp, PM, Tuohy, TM y Mosurski, KR. (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res **14**(13):5125-43.
30. Sharp, PM y Li, WH. (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res **15**(3):1281-1295.
31. Koski, LB, Morton, RA y Golding, GB. (2001). Codon bias and base composition are poor indicators of horizontally transferred genes. Mol Biol Evol **18**(3):404-412.
32. Wang, B. (2001). Limitations of compositional approach to identifying horizontally transferred genes. J Mol Evol **53**(3):244-250.
33. Ragan, MA. (2001). On surrogate methods for detecting lateral gene transfer. FEMS Microbiol Lett **201**(2):187-91.
34. Lawrence, JG y Ochman, H. (1998). Molecular archaeology of the *Escherichia coli* genome. Proc Natl Acad Sci U S A **95**(16):9413-7.
35. Yap, WH, Zhang, Z y Wang, Y. (1999). Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. J Bacteriol **181**(17):5201-9.
36. Garcia-Vallvé, S, Romeu, A y Palau, J. (2000). Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res **10**(11):1719-25.
37. Koonin, EV, Makarova, KS y Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol **55**:709-742.
38. Syvanen, M. (2002). On the occurrence of horizontal gene transfer among an arbitrarily chosen group of 26 genes. J Mol Evol **54**(2):258-66.
39. Daubin, V, Moran, NA y Ochman, H. (2003). Phylogenetics and the cohesion of bacterial genomes. Science **301**(5634):829-32.
40. Nakamura, Y, Itoh, T, Matsuda, H, et al. (2004). Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet **36**(7):760-6.
41. Doolittle, WF. (1999). Phylogenetic classification and the universal tree. Science **284**(5423):2124-2129.
42. Brown, JR, Douady, CJ, Italia, MJ, et al. (2001). Universal trees based on large combined protein sequence data sets. Nat Genet **28**(3):281-5.
43. Lerat, E, Daubin, V y Moran, NA. (2003). From Gene Trees to Organismal Phylogeny in Prokaryotes:The Case of the gamma-Proteobacteria. PLoS Biol **1**(1):E19.
44. Dykhuizen, DE y Green, L. (1991). Recombination in *Escherichia coli* and the definition of biological species. J Bacteriol **173**(22):7257-68.
45. Guttman, DS y Dykhuizen, DE. (1994). Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. Science **266**(5189):1380-3.
46. Vinuesa, P, Silva, C, Werner, D, et al. (2005). Population genetics and phylogenetic inference in bacterial molecular systematics: the roles of migration and recombination in *Bradyrhizobium* species cohesion and delineation. Mol Phylogenet Evol **34**(1):29-54.
47. Zhaxybayeva, O, Lapierre, P y Gogarten, JP. (2004). Genome mosaicism and organismal lineages. Trends Genet **20**(5):254-60.
48. Jain, R, Rivera, MC y Lake, JA. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A **96**(7):3801-6.
49. Ochman, H, Lawrence, JG y Groisman, EA. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature **405**(6784):299-304.
50. Gogarten, JP, Doolittle, WF y Lawrence, JG. (2002). Prokaryotic evolution in light of gene transfer. Mol Biol Evol **19**(12):2226-38.
51. Lawrence, JG. (2002). Gene transfer in bacteria: speciation without species? Theor Popul Biol **61**(4):449-60.
52. Kurland, CG. (2000). Something for everyone. Horizontal gene transfer in evolution. EMBO Rep **1**(2):92-5.
53. Berg, OG y Kurland, CG. (2002). Evolution of microbial genomes: sequence acquisition and loss. Mol Biol Evol **19**(12):2265-76.
54. Kurland, CG, Canback, B y Berg, OG. (2003). Horizontal gene transfer: a critical view. Proc Natl Acad Sci U S A **100**(17):9658-62.
55. Kurland, CG. (1992). Translational accuracy and the fitness of bacteria. Annu Rev Genet **26**:29-50.
56. Kimura, M (1983) *The neutral theory of molecular evolution* (Cambridge University Press, Cambrige, UK).
57. Smith, MW, Feng, DF y Doolittle, RF. (1992). Evolution by acquisition: the case for horizontal gene transfers. Trends Biochem Sci **17**(12):489-93.

58. Woese, CR y Fox, GE. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A **74**(11):5088-90.
59. Syvanen, M. (1994). Horizontal gene transfer: evidence and possible consequences. Annu Rev Genet **28**:237-61.
60. Doolittle, RF, Feng, DF, Anderson, KL, et al. (1990). A naturally occurring horizontal gene transfer from a eukaryote to a prokaryote. J Mol Evol **31**(5):383-8.
61. Martin, W, Brinkmann, H, Savonna, C, et al. (1993). Evidence for a chimeric nature of nuclear genomes: eubacterial origin of eukaryotic glyceraldehyde-3-phosphate dehydrogenase genes. Proc Natl Acad Sci U S A **90**(18):8692-6.
62. Bogusz, D, Appleby, CA, Landsmann, J, et al. (1988). Functioning haemoglobin genes in non-nodulating plants. Nature **331**(6152):178-80.
63. Smith, MW y Doolittle, RF. (1992). A comparison of evolutionary rates of the two major kinds of superoxide dismutase. J Mol Evol **34**(2):175-84.
64. Felsestein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **27**:401-410.
65. Gruskin, KD, Smith, TF y Goodman, M. (1987). Possible origin of a calmodulin gene that lacks intervening sequences. Proc Natl Acad Sci U S A **84**(6):1605-8.
66. Kemmerer, EC, Lei, M y Wu, R. (1991). Structure and molecular evolutionary analysis of a plant cytochrome c gene: surprising implications for Arabidopsis thaliana. J Mol Evol **32**(3):227-37.
67. Doolittle, RF. (1994). Convergent evolution: the need to be explicit. Trends Biochem Sci **19**(1):15-8.
68. Stewart, CB, Schilling, JW y Wilson, AC. (1987). Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature **330**(6146):401-4.
69. Hein, J. (1990). Reconstructing evolution of sequences subject to recombination using parsimony. Math Biosci **98**(2):185-200.
70. Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol **17**(6):368-76.
71. Kishino, H y Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J Mol Evol **29**(2):170-9.
72. Lake, JA. (1987). A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. Mol Biol Evol **4**(2):167-91.
73. Kidwell, MG. (1993). Lateral transfer in natural populations of eukaryotes. Annu Rev Genet **27**:235-56.
74. Clark, JB, Maddison, WP y Kidwell, MG. (1994). Phylogenetic analysis supports horizontal transfer of P transposable elements. Mol Biol Evol **11**(1):40-50.
75. Xiong, Y y Eickbush, TH. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. Embo J **9**(10):3353-62.
76. Whittam, TS, Ochman, H y Selander, RK. (1983). Multilocus genetic structure in natural populations of Escherichia coli. Proc Natl Acad Sci U S A **80**(6):1751-5.
77. Ochman, H y Selander, RK. (1984). Evidence for clonal population structure in *Escherichia coli*. Proc Natl Acad Sci U S A **81**(1):198-201.
78. Sawyer, S. (1989). Statistical tests for detecting gene conversion. Mol Biol Evol **6**(5):526-38.
79. Milkman, R y Bridges, MM. (1990). Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. Genetics **126**(3):505-17.
80. Milkman, R y Bridges, MM. (1993). Molecular evolution of the *Escherichia coli* chromosome. IV. Sequence comparisons. Genetics **133**(3):455-68.
81. Beltran, P, Musser, JM, Helmuth, R, et al. (1988). Toward a population genetic analysis of Salmonella: genetic diversity and relationships among strains of serotypes *S. choleraesuis*, *S. derby*, *S. dublin*, *S. enteritidis*, *S. heidelberg*, *S. infantis*, *S. newport*, and *S. typhimurium*. Proc Natl Acad Sci U S A **85**(20):7753-7.
82. Groisman, EA, Sturmoski, MA, Solomon, FR, et al. (1993). Molecular, functional, and evolutionary analysis of sequences specific to Salmonella. Proc Natl Acad Sci U S A **90**(3):1033-7.
83. Daubin, V, Lerat, E y Perriere, G. (2003). The source of laterally transferred genes in bacterial genomes. Genome Biol **4**(9):R57.
84. Groisman, EA, Saier, MH, Jr. y Ochman, H. (1992). Horizontal transfer of a phosphatase gene as evidence for mosaic structure of the Salmonella genome. Embo J **11**(4):1309-16.
85. Grocock, RJ y Sharp, PM. (2002). Synonymous codon usage in *Pseudomonas aeruginosa* PA01. Gene **289**(1-2):131-9.

86. Médigue, C, Rouxel, T, Vigier, P, et al. (1991). Evidence for horizontal gene transfer in *Escherichia coli* speciation. J Mol Biol **222**(4):851-856.
87. Lawrence, JG y Ochman, H. (1997). Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol **44**(4):383-397.
88. Lawrence, JG. (1999). Gene transfer, speciation, and the evolution of bacterial genomes. Curr Opin Microbiol **2**(5):519-523.
89. Ochman, H y Jones, IB. (2000). Evolutionary dynamics of full genome content in *Escherichia coli*. Embo J **19**(24):6637-6643.
90. Karlin, S, Mrázek, J, Campbell, A, et al. (2001). Characterizations of highly expressed genes of four fast-growing bacteria. J Bacteriol **183**(17):5025-40.
91. Karlin, S y Mrázek, J. (2001). Predicted highly expressed and putative alien genes of *Deinococcus radiodurans* and implications for resistance to ionizing radiation damage. Proc Natl Acad Sci U S A **98**(9):5240-5245.
92. Karlin, S, Barnett, MJ, Campbell, AM, et al. (2003). Predicting gene expression levels from codon biases in alpha-proteobacterial genomes. Proc Natl Acad Sci U S A **100**(12):7313-8.
93. Lobry, JR y Gautier, C. (1994). Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. Nucleic Acids Res **22**(15):3174-80.
94. Akashi, H y Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. Proc Natl Acad Sci U S A **99**(6):3695-700.
95. Sharp, PM y Bulmer, M. (1988). Selective differences among translation termination codons. Gene **63**(1):141-5.
96. Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. Genetics **129**(3):897-907.
97. Mathews, CK, Van-Holde, KE y Ahern, KG (2000) *Biochemistry* (Adison Wesley Longman, Oregon).
98. Brown, CM, Stockwell, PA, Trotman, CN, et al. (1990). Sequence analysis suggests that tetra-nucleotides signal the termination of protein synthesis in eukaryotes. Nucleic Acids Res **18**(21):6339-45.
99. Brown, CM, Stockwell, PA, Trotman, CN, et al. (1990). The signal for the termination of protein synthesis in procaryotes. Nucleic Acids Res **18**(8):2079-86.
100. Arkov, AL, Korolev, SV y Kisselev, LL. (1993). Termination of translation in bacteria may be modulated via specific interaction between peptide chain release factor 2 and the last peptidyl-tRNA(Ser/Phe). Nucleic Acids Res **21**(12):2891-7.
101. Eyre-Walker, A. (1996). The close proximity of *Escherichia coli* genes: consequences for stop codon and synonymous codon use. J Mol Evol **42**(2):73-8.
102. Bernardo, JO y Smith, AMF (1994) *Bayesian Theory* (John Wiley and Sons, New York).
103. Altschul, SF, Madden, TL, Schaffer, AA, et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25**(17):3389-402.
104. Karlin, S, Mrázek, J y Campbell, AM. (1998). Codon usages in different gene classes of the *Escherichia coli* genome. Mol Microbiol **29**(6):1341-55.
105. Moreno-Hagelsieb, G y Collado-Vides, J. (2002). A powerful non-homology method for the prediction of operons in prokaryotes. Bioinformatics **18 Suppl 1**:S329-36.
106. Ermolaeva, MD, White, O y Salzberg, SL. (2001). Prediction of operons in microbial genomes. Nucleic Acids Res **29**(5):1216-21.
107. Schaffer, AA, Aravind, L, Madden, TL, et al. (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res **29**(14):2994-3005.
108. Needleman, SB y Wunsch, CD. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol **48**(3):443-53.
109. Olson, SA. (2002). EMBOSS opens up sequence analysis. European Molecular Biology Open Software Suite. Brief Bioinform **3**(1):87-91.
110. Koski, LB y Golding, GB. (2001). The closest BLAST hit is often not the nearest neighbor. J Mol Evol **52**(6):540-542.
111. Thompson, JD, Higgins, DG y Gibson, TJ. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22**(22):4673-80.
112. Ward, DM. (1998). A natural species concept for prokaryotes. Curr Opin Microbiol **1**(3):271-7.
113. Nichols, R. (2001). Gene trees and species trees are not the same. Trends Ecol Evol **16**(7):358-364.

114. Moreno-Hagelsieb, G y Collado-Vides, J. (2002). Operon conservation from the point of view of *Escherichia coli*, and inference of functional interdependence of gene products from genome context. In Silico Biol **2**(2):87-95.
115. Wheeler, DL, Chappey, C, Lash, AE, et al. (2000). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res **28**(1):10-4.
116. Zeigler, DR. (2003). Gene sequences useful for predicting relatedness of whole genomes in bacteria. Int J Syst Evol Microbiol **53**(Pt 6):1893-900.
117. Endo, T, Ikeo, K y Gojobori, T. (1996). Large-scale search for genes on which positive selection may operate. Mol Biol Evol **13**(5):685-90.
118. Wilson, AC, Carlson, SS y White, TJ. (1977). Biochemical evolution. Annu Rev Biochem **46**:573-639.
119. Hurst, LD y Smith, NG. (1999). Do essential genes evolve slowly? Curr Biol **9**(14):747-750.
120. Hirsh, AE y Fraser, HB. (2001). Protein dispensability and rate of evolution. Nature **411**(6841):1046-9.
121. Pal, C, Papp, B y Hurst, LD. (2003). Genomic function: Rate of evolution and gene dispensability. Nature **421**(6922):496-7; discussion 497-8.
122. Liu, Y, Harrison, PM, Kunin, V, et al. (2004). Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. Genome Biol **5**(9):R64.
123. Taoka, M, Yamauchi, Y, Shinkawa, T, et al. (2004). Only a small subset of the horizontally transferred chromosomal genes in *Escherichia coli* are translated into proteins. Mol Cell Proteomics **3**(8):780-7.
124. Jain, R, Rivera, MC, Moore, JE, et al. (2003). Horizontal gene transfer accelerates genome innovation and evolution. Mol Biol Evol **20**(10):1598-602.
125. Datsenko, KA y Wanner, BL. (2000). One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. Proc Natl Acad Sci U S A **97**(12):6640-5.
126. Hand, D y Keming, Y. (2001). Idiot's Bayes-Not So Stupid After All? International Statistical Review **69**(3):385-398.
127. Fiers, W, Contreras, R, Duerinck, F, et al. (1975). A-protein gene of bacteriophage MS2. Nature **256**(5515):273-8.
128. Air, GM, Blackburn, EH, Coulson, AR, et al. (1976). Gene F of bacteriophage phiX174. Correlation of nucleotide sequences from the DNA and amino acid sequences from the gene product. J Mol Biol **107**(4):445-58.
129. Efstratiadis, A, Kafatos, FC y Maniatis, T. (1977). The primary structure of rabbit beta-globin mRNA as determined from cloned DNA. Cell **10**(4):571-85.
130. Clarke, B. (1970). Darwinian evolution of proteins. Science **168**(934):1009-11.
131. Bennetzen, JL y Hall, BD. (1982). Codon selection in yeast. J Biol Chem **257**(6):3026-31.
132. Gouy, M y Gautier, C. (1982). Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res **10**(22):7055-74.
133. Grosjean, H y Fiers, W. (1982). Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene **18**(3):199-209.
134. Konigsberg, W y Godson, GN. (1983). Evidence for use of rare codons in the *dnaG* gene and other regulatory genes of *Escherichia coli*. Proc Natl Acad Sci U S A **80**(3):687-91.
135. Bennett, AD y Shaw, WV. (1983). Resistance to fusidic acid in *Escherichia coli* mediated by the type I variant of chloramphenicol acetyltransferase. A plasmid-encoded mechanism involving antibiotic binding. Biochem J **215**(1):29-38.
136. Sharp, PM y Li, WH. (1986). Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. Nucleic Acids Res **14**(19):7737-49.
137. Varenne, S, Buc, J, Lloubes, R, et al. (1984). Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. J Mol Biol **180**(3):549-76.
138. Burns, DM y Beacham, IR. (1985). Rare codons in *E. coli* and *S. typhimurium* signal sequences. FEBS Lett **189**(2):318-24.
139. Bonekamp, F, Andersen, HD, Christensen, T, et al. (1985). Codon-defined ribosomal pausing in *Escherichia coli* detected by using the *pyrE* attenuator to probe the coupling between transcription and translation. Nucleic Acids Res **13**(11):4113-23.
140. Bonekamp, F, Dalboge, H, Christensen, T, et al. (1989). Translation rates of individual codons are not correlated with tRNA abundances or with frequencies of utilization in *Escherichia coli*. J Bacteriol **171**(11):5812-6.
141. Pedersen, S. (1984). *Escherichia coli* ribosomes translate in vivo with variable rate. Embo J **3**(12):2895-8.

142. Holm, L. (1986). Codon usage and gene expression. Nucleic Acids Res **14**(7):3075-87.
143. Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol **2**(1):13-34.
144. Fitch, DH y Strausbaugh, LD. (1993). Low codon bias and high rates of synonymous substitution in *Drosophila hydei* and *D. melanogaster* histone genes. Mol Biol Evol **10**(2):397-413.
145. Akashi, H. (1994). Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics **136**(3):927-35.
146. Gribskov, M, Devereux, J y Burgess, RR. (1984). The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. Nucleic Acids Res **12**(1 Pt 2):539-49.
147. McLachlan, AD, Staden, R y Boswell, DR. (1984). A method for measuring the non-random bias of a codon usage table. Nucleic Acids Res **12**(24):9567-75.
148. Karlin, S y Mrázek, J. (2000). Predicted highly expressed genes of diverse prokaryotic genomes. J Bacteriol **182**(18):5238-50.
149. Ma, J, Campbell, A y Karlin, S. (2002). Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. J Bacteriol **184**(20):5733-45.
150. Sharp, PM y Matassi, G. (1994). Codon usage and genome evolution. Curr Opin Genet Dev **4**(6):851-860.
151. Lafay, B, Lloyd, AT, McLean, MJ, et al. (1999). Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucleic Acids Res **27**(7):1642-1649.
152. Lafay, B, Atherton, JC y Sharp, PM. (2000). Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. Microbiology **146**(Pt 4):851-860.
153. Dong, H, Nilsson, L y Kurland, CG. (1996). Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. J Mol Biol **260**(5):649-63.
154. Pedersen, S, Bloch, PL, Reeh, S, et al. (1978). Patterns of protein synthesis in *E. coli*: a catalog of the amount of 140 individual proteins at different growth rates. Cell **14**(1):179-90.
155. VanBogelen, RA, Sankar, P, Clark, RL, et al. (1992). The gene-protein database of *Escherichia coli*: edition 5. Electrophoresis **13**(12):1014-54.
156. Garcia, GM, Mar, PK, Mullin, DA, et al. (1986). The *E. coli dnaY* gene encodes an arginine transfer RNA. Cell **45**(3):453-9.
157. Komine, Y, Adachi, T, Inokuchi, H, et al. (1990). Genomic organization and physical mapping of the transfer RNA genes in *Escherichia coli* K12. J Mol Biol **212**(4):579-98.
158. Saxena, P y Walker, JR. (1992). Expression of *argU*, the *Escherichia coli* gene coding for a rare arginine tRNA. J Bacteriol **174**(6):1956-64.
159. Björk, GR (1995) in *In tRNA: Structure, Biosynthesis, and Function*, eds. Söll, D y RajBhandary, UL (American Society for Microbiology, Washington, DC.), pp. 165-205.
160. Bremer, H y Dennis, PP (1987) in *Escherichia coli and Salmonella typhimurium cellular and molecular biology*, ed. Ingraham, JL, Low, K. B., Magasanik, B., Schaechter, M., Umbarger, H. E. (American Society for Microbiology, Washington, DC.), Vol. 2, pp. 1527-1542.
161. Donachie, WD y Robinson, AC (1987) in *Escherichia coli and Salmonella typhimurium cellular and molecular biology*, ed. Ingraham, JL, Low, K. B., Magasanik, B., Schaechter, M., Umbarger, H. E. (American society for microbiology, Washington, DC.), Vol. 2, pp. 1578-1593.
162. Carbone, A, Zinovyev, A y Kepes, F. (2003). Codon adaptation index as a measure of dominating codon bias. Bioinformatics **19**(16):2005-15.
163. Ramakrishnan, V y White, SW. (1998). Ribosomal protein structures: insights into the architecture, machinery and evolution of the ribosome. Trends Biochem Sci **23**(6):208-12.
164. Smith, NG y Eyre-Walker, A. (2001). Why Are Translationally Sub-Optimal Synonymous Codons Used in *Escherichia coli*? J Mol Evol **53**(3):225-236.
165. Yamao, F, Andachi, Y, Muto, A, et al. (1991). Levels of tRNAs in bacterial cells as affected by amino acid usage in proteins. Nucleic Acids Res **19**(22):6119-22.
166. Christen, JA, Torres, JL y Barrera, J. (1998). A statistical feature of genetic sequences. Biometrical journal **40**(7):855-863.
167. Singer, GA y Hickey, DA. (2000). Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol Biol Evol **17**(11):1581-8.