Volumen 10
XXVI Verano de la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

Análisis filogenético y estructural de factores transcripcionales de *Escherichia coli* de función desconocida

Alondra Aguillón Bárcenas[1], Isabel Duarte Velázquez[1], Fátima Tornero Gutiérrez[1], Eugenia Cordero Loreto[2], Naurú Idalia Vargas Maya[1], Felipe Padilla Vaca[1], Bernardo Franco[1]
[1]División de Ciencias Naturales y Exactas, Universidad de Guanajuato.
[2]Colegio del Nivel Medio Superior, Universidad de Guanajuato

Resumen

Todos los seres vivos tienen mecanismos moleculares que les permiten percibir lo que ocurre en el exterior de sus tejidos, células o individuos. Las bacterias no son la excepción, tienen el enorme reto de contender con condiciones cambiantes del medio ambiente y que requieren responder de manera precisa y rápida para poder adaptarse al medio ambiente y sobrevivir. El principal mecanismo de respuesta caracterizado en bacterias es modificar los patrones de expresión génica en respuesta a estímulos ambientales, esto se logra por la activación o represión diferencial de genes.

Los factores transcripcionales son proteínas que regulan el uso de un número limitado de RNA polimerasas celulares. Estos factores permiten reconocer secuencias específicas en el DNA (secuencias blanco) que están cerca del promotor de genes necesarios para dar una respuesta eficiente ante un estímulo ambiental.

En el genoma de la bacteria entérica Escherichia coli, se han identificado 58 marcos de lectura con una homología suficiente para identificarlos como factores transcripcionales, y que su función biológica sigue sin ser caracterizada experimentalmente. Algunos tienen aparentemente un origen viral (bacteriófagos). Las herramientas bioinformáticas permiten hacer análisis que pueden derivar en hipótesis comprobables del origen, evolución y función de proteínas de función desconocida. Por tanto, en este trabajo se pretende analizar 58 factores transcripcionales de función desconocida usando herramientas computacionales y plantear una hipótesis evolutiva de estos y su distribución en otras bacterias. Los resultados obtenidos sugieren que, de los 58 factores, una proporción importante de estos tienen secuencias comunes que son independientes de la familia a la que potencialmente se les puede asignar. El análisis de su posición genómica sugiere que muchos de estos fueron adquiridos por algún mecanismo de transferencia horizontal y que esto pudo ser parte de la evolución tardía del genoma de E. coli, permitiendo tener un grupo de factores transcripcionales que en alguna condición especial serán funcionales. En general, tenemos un grupo de factores que comparten características de secuencia, posible estructura y quizá puedan tener funciones en el curso evolutivo de E. coli.

Palabras clave: Factores transcripcionales; Escherichia coli; Bioinformática; análisis de función; Filogenia.

Introducción

Escherichia coli es sin duda uno de los organismos más estudiados y mejor caracterizados a nivel molecular, celular y ambiental, especialmente su papel como un comensal común en el intestino de los vertebrados. Con la publicación del genoma de este microorganismo, se pudo caracterizar muchos aspectos de esta, principalmente sobre su metabolismo especialmente el descubrimiento de un operón para la degradación de compuestos aromáticos, una similitud enorme con Salmonella typhimurium en cuanto a su arquitectura genómica y especialmente en los genes que codifican para flagelos, (Blattner *et al.*, 1997). El genoma de E. coli está compuesto por 4,639,221 pares de bases y hasta el momento se sabe que codifica para 4288 proteínas, de las cuales el 38% no se les ha podido asignar una función clara o bien, son pseudo genes que potencialmente podrían tener una función en una condición determinada.

Sin duda, el análisis de genomas no es una historia que termina una vez que se obtiene el análisis inicial, hay muchos aspectos que se deben estar revisando continuamente. Uno muy importante es reconstruir la historia evolutiva de las bacterias con énfasis en la descendencia y las relaciones evolutivas como lo son los genes parálogos, aquellos que descienden de genes de un progenitor y que se han duplicado generando divergencia previa a un evento de especiación (Riley y Labedan, 1997). Riley y Labedan (1997) descubrieron que en E. coli existe un número importante de parálogos en el genoma de este microorganismo, la mayoría de estos con una estructura modular, es decir, secciones de estas proteínas que contienen elementos en común. Un aspecto sorprendente de la estructura modular de E. coli es que, de los 1,404 módulos identificados, estos provienen de tan solo 352 módulos funcionales ancestrales (Riley y Labedan, 1997). Uno de los retos importantes es poder rastrear la historia evolutiva de los parálogos simplemente por la comparación de las secuencias. Para esto, otros elementos genómicos se deben tener en cuenta, como son los elementos móviles, los profagos (bacteriófagos insertados en el genoma) y secuencias repetidas.

Sin duda el estudio de la dinámica de los genomas en relación con la patogenicidad es uno de los temas centrales en genómica funcional y el estudio de cepas y especies del género Escherichia, sin embargo, poca atención se ha puesto a elementos genómicos como son los genes de función desconocida, principalmente por la ausencia de evidencia experimental que sugiera cuál puede ser su papel fisiológico. Muchas de las proteínas que hasta el momento han sido caracterizadas extensamente son proteínas involucradas en procesos metabólicos, de transporte, señalización y regulación, pero es mucho más complicado rastrear la función molecular de proteínas de función desconocida, especialmente si las mutantes de estas no presentan un fenotipo evidente bajo condiciones experimentales del laboratorio (Baba *et al.*, 2006, Yamamoto *et al.*, 2009). La colección de mutantes sencillas en cada uno de los genes conocida como la colección Keio, ha demostrado que todavía existen muchos genes que pueden ser mutados, sin un fenotipo aparente, pero que en condiciones muy particulares pueden tener relevancia, como son condiciones de estrés. Afortunadamente se cuenta con un repositorio de información de E. coli como es EcoCyc (Kesler *et al.*, 2017), el cual contiene mucha información recopilada de la literatura que ayuda a comprender el papel de cada gen de los 4,288 genes. Hasta ahora se han logrado obtener mutantes de 3985 genes, siendo el resto letales. Es importante señalar que de los 303 genes que no se han podido obtener mutantes, 37 son de función desconocida (Baba *et al.*, 2006), dentro de los cuales, ninguno es un factor transcripcional.

Por lo anterior, en el presente trabajo proporcionamos algunas observaciones bioinformáticas que sugieren que los factores transcripcionales de función desconocida pertenecen a un grupo de genes que provienen de eventos de transferencia horizontal y que pueden permanecer como amortiguadores de eventos mutacionales en factores transcripcionales de función esencial en la célula. Eventualmente pueden presentar un cambio de especificidad y poder expresarse en condiciones desfavorables o bien, en caso de pérdida de función de factores involucrados en la regulación de genes de estrés.

Objetivos

General:
Analizar 58 factores transcripcionales de E. coli para evaluar su relación filogenética.
Particulares:

- Analizar los 58 factores transcripcionales de función desconocida de E. coli a nivel de secuencia y estructura para establecer su relación filogenética
- Establecer la sintenia de los diferentes factores transcripcionales para poder establecer una hipótesis sobre la distribución de estos entre otras bacterias.

Hipótesis

Algunos de los factores transcripcionales de función desconocida tendrán elementos que sugieren ser parte de regiones genómicas ya sea en rearreglos constantes o bien, de origen viral (bacteriófagos).
Resultados:
La localización genómica de los factores transcripcionales es el primer análisis que se exploró en el presente trabajo para poder realizar un mapa fiel de su localización y tratar de relacionar su posición en el genoma de E. coli con la de elementos genéticos móviles, transposones y profagos. Para esto, se realizó un mapa físico usando la herramienta CGview (Grant y Stothard, 2008), resaltando las regiones del genoma que contienen los 58 factores transcripcionales de función desconocida.

En la Figura 1 se muestra el mapa genómico de E. coli K-12 cepa MG1655 en la que se puede apreciar el sesgo del contenido de G+C en el genoma correlaciona con la posición de dos grandes cúmulos de factores transcripcionales de función desconocida, codificados tanto en la cadena positiva como en la negativa (anillo externo e interno respectivamente), además de la presencia de varios de estos factores transcripcionales localizados con secuencias de profagos crípticos, siendo una región donde encontramos varios de estos factores transcripcionales en la región de 0.5 y 1 Mpb donde además hay una zona concentrada de elementos móviles. El mapa además señala la posición de secuencias repetidas en la vecindad de muchos de estos, especialmente en la zona de los 3.5 y 4 Mbp. De manera general, algunos de estos factores transcripcionales corresponden a zonas donde tenemos muchos elementos que suelen estar asociados a movilidad genómica, especialmente a secuencias REP (Repetitive Extragenic Palindrome) por ser sitios frecuentes de recombinación o transposición, aunque su posible función también esté en estabilizar el RNA mensajero de genes, o bien como blancos para la unión de la enzima DNA girasa (Gilson *et al.*, 1984). Las secuencias REP tienen como característica que forman estructuras tallo-asa y potencialmente son sitios que hayan resultado de recombinación dada su naturaleza palindrómica (https://ecocyc.org/ECOLI/NEW-IMAGE?type=ECOCYC-CLASS&object=REP-Elements).

Volumen 10
XXVI Verano de la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

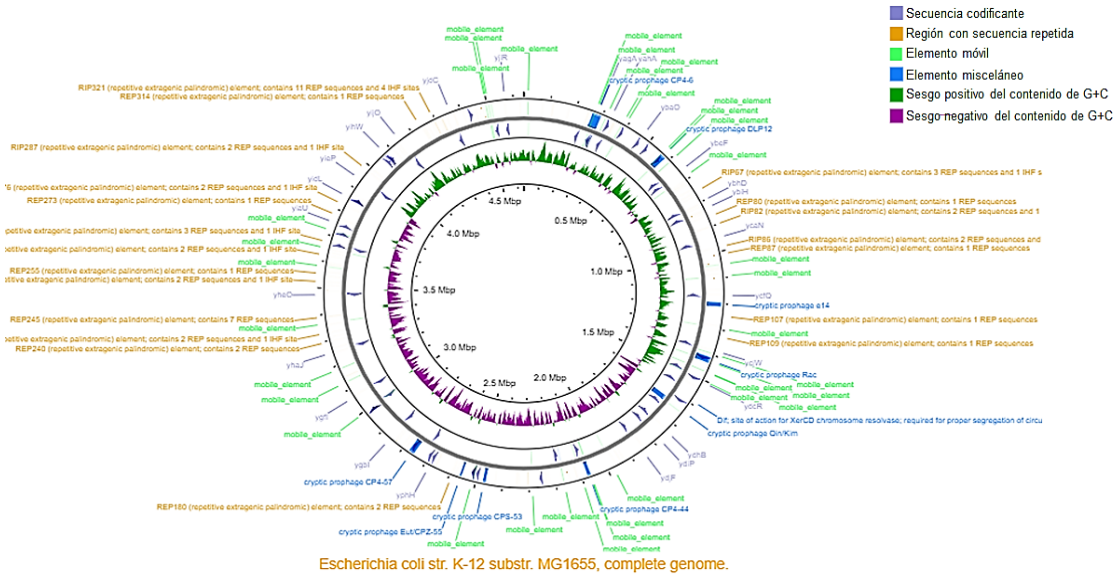


Figura 1. Mapa físico del genoma de E. coli K-12 cepa MG1655, indicando únicamente las secuencias codificantes de los factores transcripcionales de función desconocida y otros elementos genómicos, como son regiones con secuencias repetidas, elementos móviles, elementos misceláneos como son orígenes de replicación y profagos, se indica el sesgo del contenido de G+C tanto positivo como negativo (GCView. *et al*., 2011)

El análisis del Codon Adaptation Index (CAI), que hace referencia al tipo de codones preferentemente empleados por un organismo para traducir sus proteínas, también es un indicador sugerente sobre la capacidad de un mRNA para ser traducido e incluso en los niveles de mRNA e incluso la estabilidad del mRNA en E. coli (Boël *et al.*, 2016) e incluso en humanos, así como las diferencias de traducción en mitocondria (Lavner y Kotlar, 2005). En este sentido, se evaluó la preferencia del uso de codones y el contenido de G+C en la tercera posición de cada codón de las secuencias codificantes de los 58 factores de transcripción analizados en el presente trabajo con miras a evaluar su posible relación con su expresión y potencialmente su función. En la Figura 2 se muestran los resultados.

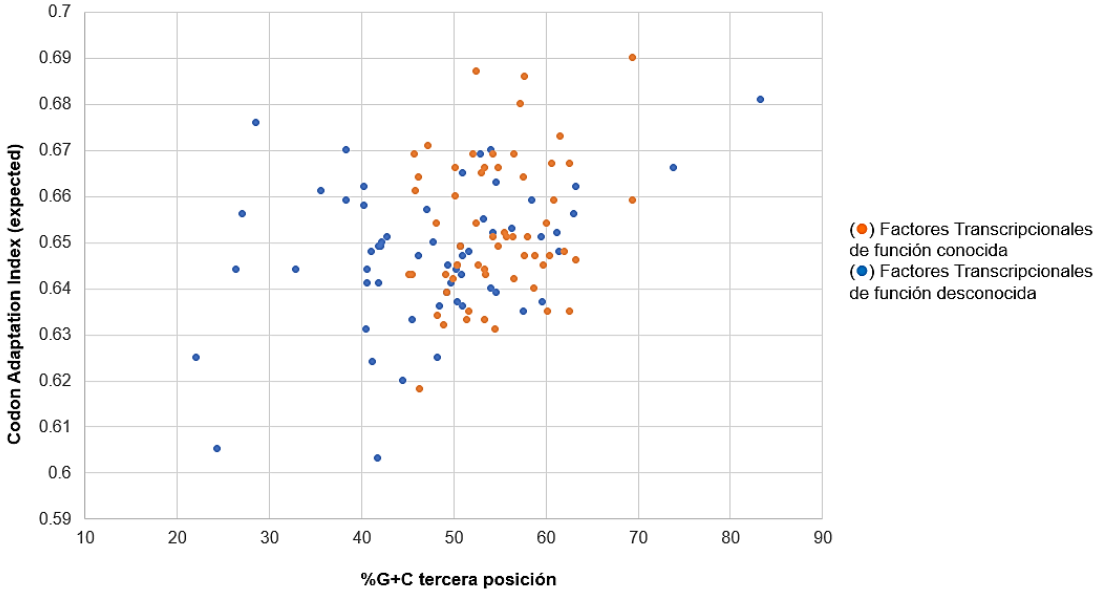


Figura 2. Análisis del uso de codones esperado para cada factor transcripcional y su correlación con la variación en el contenido de G+C en la tercera posición de cada codón. Se muestra el resultado comparativo del contenido de G+C en la tercera posición de los codones de 58 factores transcripcionales de función desconocida (en azul) y 58 de función conocida (en anaranjado) contra el CAI esperado (eCAI) de cada uno.

Los resultados sugieren que los factores transcripcionales de función conocida tienen un valor de G+C y CAI que representan un cúmulo dentro de parámetros semejantes. Sin embargo, los factores transcripcionales de función desconocida tienen una mayor dispersión en cuanto a estos dos parámetros, lo que sugiere tanto un origen externo a E. coli y también, que su expresión puede ser mucho menor que los factores de transcripción de función conocida.

Con base en el resultado anterior, decidimos evaluar la secuencia promotora de los genes codificantes para factores de transcripción de función desconocida. Una característica que tienen los elementos que han sido adquiridos por transferencia horizontal, es que las secuencias promotoras tienen un mayor contenido de A+T que el resto de los promotores bacterianos, esto debido a la formación de islas con bajo contenido de G+C que resulta en secuencias reconocidas por factores como H-NS o proteína tipo histona, que se asocia con el silenciamiento de genes (Purtov *et al.*, 2014) y más recientemente, experimentalmente se ha demostrado que estas secuencias potencialmente pueden acumular mutaciones y volverse regiones activas, lejos del silenciamiento mediado por H-NS o la proteína similar a histonas (Bykov *et al.*, 2020) buscamos secuencias ricas en A+T en los promotores de los genes de TF desconocidos y encontramos dos secuencias de alta representatividad en algunos de los promotores de los factores de función desconocida.

En la Figura 4 se muestra el análisis hecho buscando motivos más frecuentemente representados en las secuencias analizadas, encontramos dos con una alta puntuación (Figura 4 panel A y Panel C), de los cuales, resalta la presencia de un alto contenido de A+T. La posición relativa de estas secuencias indica su cercanía al inicio de la transcripción de los genes analizados, pero solamente algunos de los promotores tienen estos elementos. Para el motivo mostrado en la Figura 4 panel A corresponde a los factores de transcripción: ygeK, yiaU, yqeI, yfeD, yfeC, ybdO, ybeF, ybiH, ycfQ, ydcI, ydiP, ydjF, yfiE, yfjR yieP, yjiR. Para el motivo mostrado en la Figura 4 panel C, solamente es encontrado en los genes: yqeH, ydcI, yqeI, ycfQ, yfhH, ybdO, yddM, ybhD, yfiE, yfjR, ygfI, yheO y yiaG. Mediante la herramienta Tomtom (Gupta *et al.*, 2007) se identificó que el motivo mostrado en la Figura 4 panel A tiene un score elevado con la secuencia que reconoce el factor transcripcional modE, el cual regula la expresión de los genes de enzimas y funciones relacionadas con el molibdeno (EcoCyc G6395) y nagC, el cual regula la biosíntesis de amino azucares como la D-glucosamina y N-acetil glucosamina (EcoCyc EG10636). El motivo mostrado en la Figura 4 panel C tiene similitud con el sitio de reconocimiento de fur, que es un factor transcripcional que requiere hierro (Fe^{2+}) y es represor de la expresión de proteínas de la membrana externa incluyendo el operón de transporte de hierro, controlando la homeostasis de hierro (EcoCyc EG10359). Estos resultados de manera colectiva sugieren que se trata de genes con potencial de expresión y regulación con procesos celulares, lo que potencialmente puede ser un paso previo a ser genes activos o que pueden suplir su función.

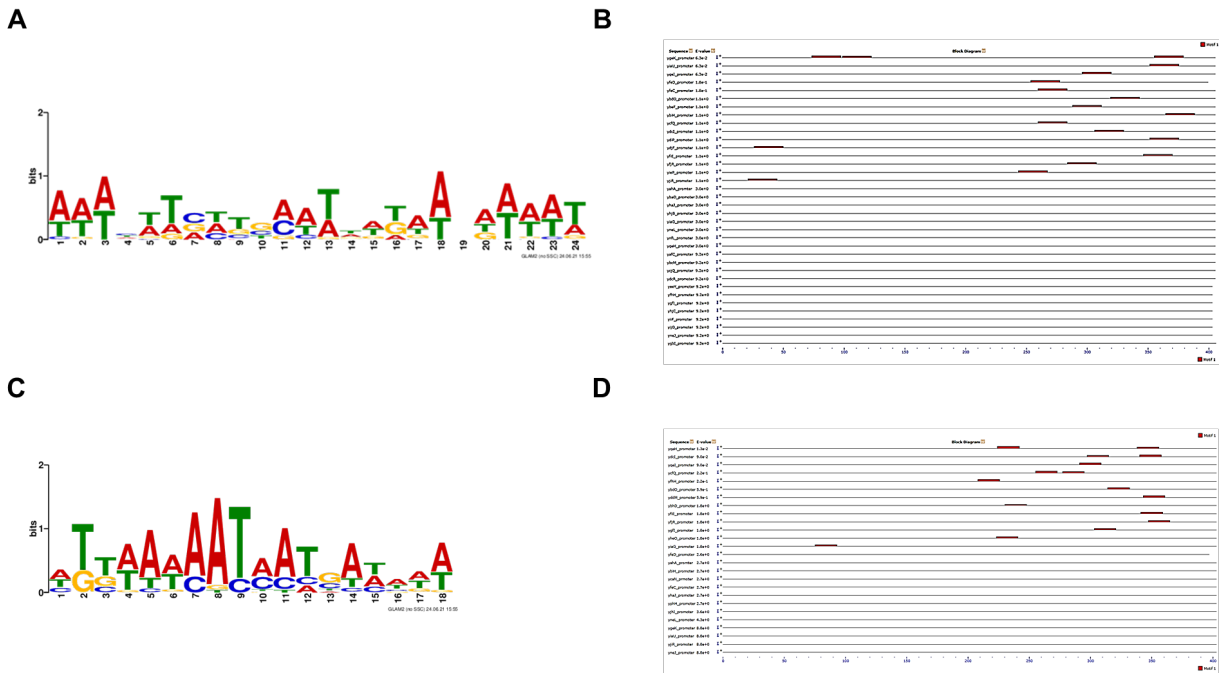


Figura 4. Análisis se secuencias ricas en A+T en las regiones promotoras de los 58 factores transcripcionales de función desconocida. El análisis muestra dos secuencias abundantes ricas en A o T y que solamente está presente en un grupo de

factores transcripcionales. Panel A, primer motivo de secuencia representada con su posición relativa en los promotores analizados (Panel B). Panel C, segundo motivo de secuencia representada y su posición relativa (panel D).

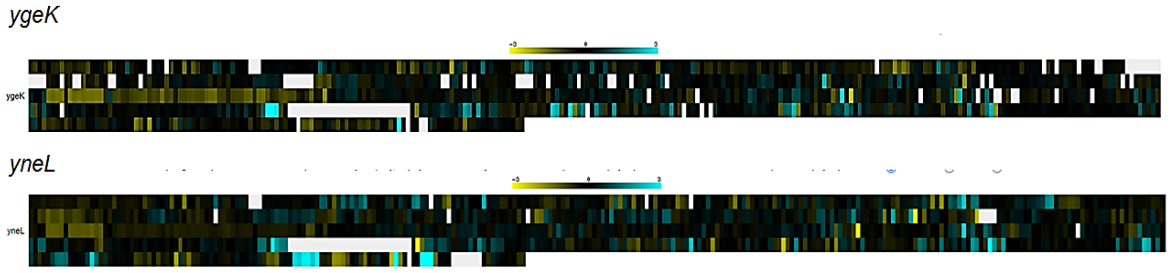
Usando una herramienta adicional para analizar la presencia de promotores (de Jong *et al.*, 2012), se encontraron los promotores con una buena puntuación para los genes mostrados en la Tabla 1.

Tabla 1. Promotores predichos con PePPER en los genes codificantes para los factores transcripcionales de función desconocida.

Nombre	Posición*	Score	Secuencia
yfeD	52	8.32228499658	TTTACGTACCAAGTTTGCTGGGTGCAAAAT
yneJ	342	10.3649024511	TTTACTCTTGCTTTAAAATGAATAATAT
yiaU	270	9.80366369168	TTAACATGTCCGGTATTCCATTTTAAAAT
yfeC	58	8.32228499658	TTTACGTACCAAGTTTGCTGGGTGCAAAAT
ybeF	359	10.4079128701	TTTAAATATTATTTTTCCATGAATAAAAT

*Indica la posición relativa del promotor con el inicio del ORF, el número más cercano a 400 indica que está cerca del ATG del ORF.

Como se puede observar, los promotores predichos son similares entre sí y ricos en A+T. Esto coincide con lo reportado previamente con genes de origen por transferencia horizontal (Daubin y Ochman, 2004, Huang *et al.*, 2012). Estas secuencias potencialmente pueden acumular mutaciones y hacer estos genes transcripcionalmente más activos, posiblemente para suplir la pérdida de función de un gen de función conocida (Bykov *et al.*, 2020).
Por otro lado, se hizo un análisis de dos supuestos pseudo genes que se encontraron entre los 58 factores transcripcionales analizados previamente, estos son: YgeK e YneL (ver Figura 5). Cabe mencionar, que si bien, se tienen anotados como pseudo genes, hay evidencia que demuestra que en ciertas condiciones se pueden expresar, por lo que tienen los suficientes elementos para poder considerarse como genes, en otras palabras, si un gen no tiene los elementos necesarios para ser un gen, este puede potencialmente tenerlos, sin embargo, esto no se muestra en los análisis bioinformáticos tradicionales. La base de datos GenExpDB (2015, https://genexpdb.okstate.edu/databases/genexpdb/), contiene evidencia de expresión de 216 experimentos de expresión por microarreglos (véase Figura 5). En este mismo sentido, la literatura reporta que YgeK (Yamamoto, *et al.*, 2005) parece ser parte de un sistema de dos componentes (una proteína sensora de tipo histidin cinasa, haciendo pareja con este factor transcripcional como un regulador de respuesta formando un sistema de transducción de señales del cual se desconoce la señal que lo activa. Lo más interesante es que YgeK, es único, porque carece del dominio receptor, que suele contener el aminoácido aspártico que es altamente conservado en estos sistemas y que es fosforilado por la histidin cinasa. Lo más sorprendente es que dos histidin cinasas, BarA y UhpB, fueron capaces de fosforilar a YgeK, lo que sugiere una fosforilación novedosa distinta de la típica, la cual es una transferencia del grupo fosfato de una histidina a un aspártico.



Volumen 10
XXVI Verano de la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

Figura 5. Perfiles de expresión de los dos factores transcripcionales de función desconocida anotados como pseudo genes. Se empleó la base de datos GenExpDB (*https://genexpdb.okstate.edu/databases/genexpdb/*) para verificar su perfil de expresión. Se muestran los datos para estos dos genes de 216 experimentos de expresión por microarreglos. El heatmap muestra desde represión (-3) hasta sobre expresión (3).

Con la finalidad de observar que dominios de los diferentes factores de transcripción permanecían conservados entre sí. Primero se utilizó el un árbol filogenético guía o guide tree obtenido del alineamiento de las 58 secuencias (llevado a cabo con la herramienta MUSCLE que cuenta con un acceso libre a través de internet (Edgar, 2004) para obtener 13 diferentes grupos de comparación (siguiendo el criterio de similitud en la secuencia entre los factores) y evitando el sesgo con la secuencia del dominio de unión a DNA (ver Figura 6) Con los grupos formados se llevó a cabo un análisis de alineamiento estructural de estos mediante el uso de la herramienta RaptorX (Wang, *et al.*, 2013; Wang *et al.*, 2011) en los que se comparó los factores de cada grupo, con el ancestro en común.

Los resultados arrojados mostraron la presencia de una porción continua y conservada de un dominio en común entre los factores y el ancestro en común. El grupo 1 mostro ser el más conservado y tener un dominio mayor en común con el ancestro; el grupo 2, 3, 10, 11 y 12 presentaron en específico el dominio involucrado en la unión de DNA presente en el ancestro; el grupo 4, 9 y 13 por su parte conservan un dominio con mayor similitud con el factor transcripcional más ancestral que cualquier otro grupo; finalmente el grupo 6 mostro ser el que presento menos similitud con el ancestro en común).

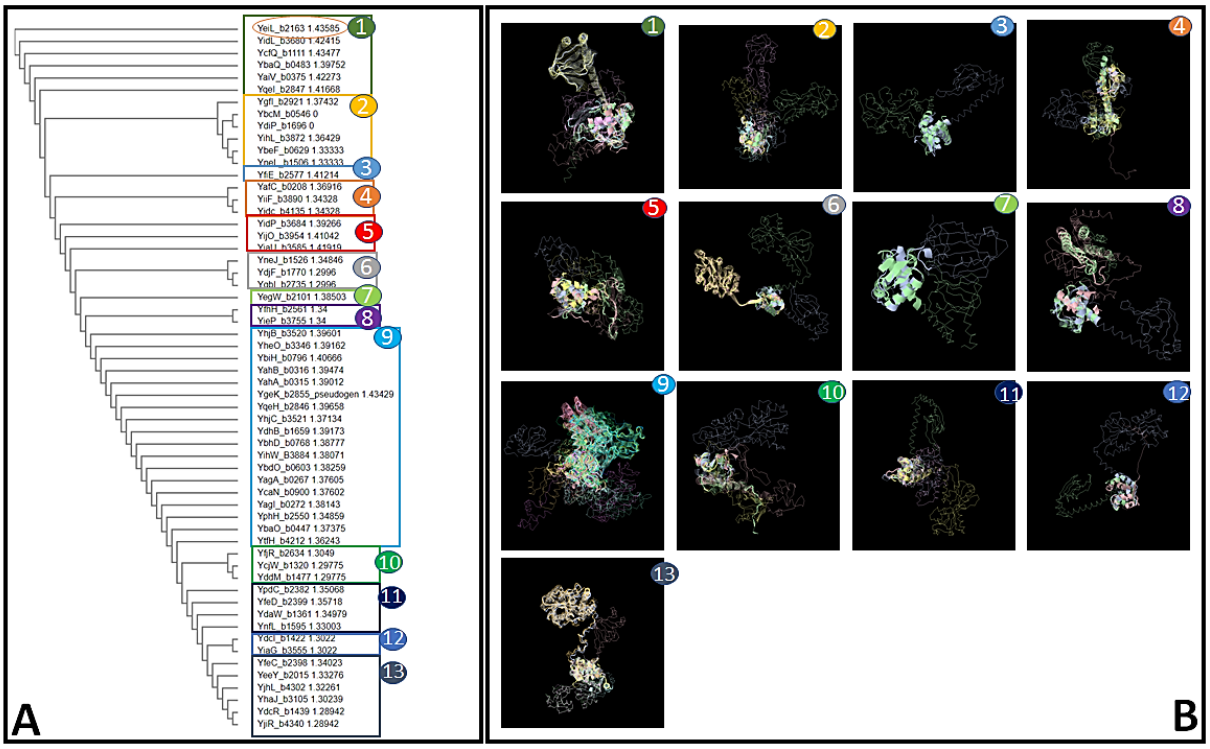


Figura 6. 2 Análisis filogenético y estructural de los 58 factores transcripcionales de función desconocida. En A) Guide tree obtenido del proceso de alineamiento llevado a cabo con la herramienta MUSCLE, se observa a manera de recuadros de colores y números los diferentes grupos (en total 13) en los que se dividió para su análisis; en la parte superior del árbol se encuentra encerrado con un círculo el ancestro en común de los factores analizados. B) Resultados del análisis de alineamiento estructural (llevado a cabo con la herramienta RaptorX); se observa en cada imagen los dominios conservados entre los diferentes factores comparados.

Volumen 10
XXVI Verano de la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

El análisis general de las secuencias muestra rasgos de conservación y con respecto a una secuencia más ancestral, esto aunado con la presencia de múltiples elementos asociados con eventos de transferencia horizontal y una clara variación en el contenido de G+C, hacen de estos factores candidatos a ser analizados con más detalle para determinar cómo y de dónde fueron transferidos. Para esto, se realizó con la ayuda de la herramienta GLAM2 se hizo un análisis comparativo de estos 58 factores transcripcionales (ver Figura 7) con la final de encontrar los motivos estructurales más conservados entre estos. Los resultados obtenidos mostraron la presencia de al menos una secuencia altamente representada entre los 58 factores (ver panel A de la Figura 7), además, una parte de los factores transcripcionales presentan esta secuencia sobrerrepresentada en el extremo animo terminal mientras que otras lo tienen cercana al carboxilo terminal.

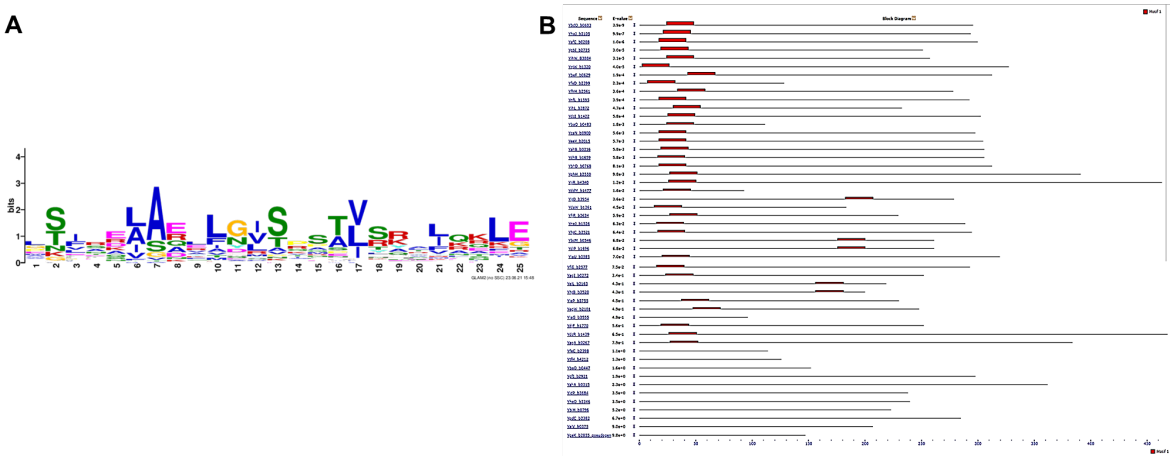


Figura 7. Análisis comparativo de los 58 factores transcripcionales de E. coli. Se hizo un análisis usando la herramienta GLAM2 (Frith *et al*., 2008) para descubrir motivos de secuencias más representadas en los 58 factores transcripcionales (Panel A). Posteriormente, se hizo un mapa físico de estas secuencias más representadas en todos los factores transcripcionales (Panel B) y se muestran del amino al carboxilo terminal (MAST, *et al*., 1998).

Volumen 10
XXVI Verano de la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

Todos los factores transcripcionales se analizaron por STRING (Szklarczyk, *et al.*, 2019). Con los resultados obtenidos se deduce que a pesar de que estos genes no han sido caracterizados, pueden estar potencialmente relacionados a las funciones que tienen sus parejas funcionales analizadas por STRING. Asimismo, se analizó si las parejas funcionales aparecían dentro de la sintenia de los genes en cuestión, lo que arrojó dos resultados, si lo que se observa en STRING correlaciona con los vecinos, entonces se propone que estos genes fueron transferidos juntos, es decir, una relación evolutiva; de no ser así, fueron transferidos junto con esos genes de otros organismos, es decir, que a lo largo del tiempo han ido ganando la función de regular genes del hospedero. Del análisis de los 58 factores transcripcionales, solo 15 mostraron una correlación con la sintenia y con al menos una de sus parejas funcionales, dichos resultados se muestran en la Tabla 2.

Posibles parejas funcionales de los factores transcripcionales de función desconocida de E. coli		
Factor	Parejas funcionales*	Funciones importantes **
YdjF	YdjK, ydjE	Proteína de transporte (supuesta, metabolitos de membrana interna)
	YdjJ	Proteína no caracterizada
	YdjG	Metilglioxal reductasa especifica de NADH
YfiE	eamB	Transporte (salida de cisteína/o-acetilserina)
YieP	HsrA	Transporte (tipo MFS)
YjiR	Rsd	Regulador (sigma D)
yfjR	YfjP	Proteína no caracterizada
	yfjQ	Relacionado a fagos o profagos
YiaU	yiaT	Función estructural (Proteína de la membrana externa)
yqeI	ygeH	Proteína no caracterizada
	yqeJ; yqeH	Función desconocida
yfeD	yfeC	Función desconocida
yfeC	yfeD	Función desconocida (pspD proteína de choque de fagos)
ybdO	ybdN	Función desconocida
ybeF	lipA; tatE	(Lipoil sintasa; Proteína translocasa)
ybiH	ybhF, ybhG	Función celular (Transportador ABC de ATP; Función estructural proteína de membrana)
	ybhS; YbhR	Función desconocida
yafC	yafE	Metabolismo de la biotina
yeiL	rihB	Función enzimática (Ribonucleósido hidrolasa específica de pirimidina)
yjdC	cutA	Resistencia a metales pesados
	dsbD	Proteína de intercambio de disulfuro

Tabla 2. Se muestran las parejas funcionales predichas para los 58 factores transcripcionales analizados en este trabajo. En la columna 1 se enlistan aquellos factores con parejas funcionales que además guardan sinténia. En la columna 2 se indican las parejas funcionales mediante un análisis usando la herramienta STRING, así como sus funciones importantes identificadas (columna 3).
* Aquellos genes sinténicos de los factores transcripcionales en cuestión se encuentran resaltados en color verde.
** Las funciones importantes están ordenadas conforme a las parejas funcionales.

Discusión

El análisis a nivel genómico es fundamental para comprender la relación de la posición de un gen y sus niveles de expresión. Recientemente, Scholz y colaboradores (2019) reportaron que el genoma de la bacteria modelo E. coli presenta un "paisaje" de expresión de genes diferencial en todo el cromosoma. Mediante el uso de una estrategia de integrar en puntos definidos del genoma secuencias reporteras, se logró determinar los niveles de expresión en todo el genoma, ando un mapa muy definido que en la región de 3.5 a 4 Mpb son las más activas transcripcionalmente. La región que comprende entre 0.5 y 2.5 Mpb son las menos activas, que además se localizan varios profagos insertados, junto con una alta densidad de factores transcripcionales de función desconocida (Figura 8).

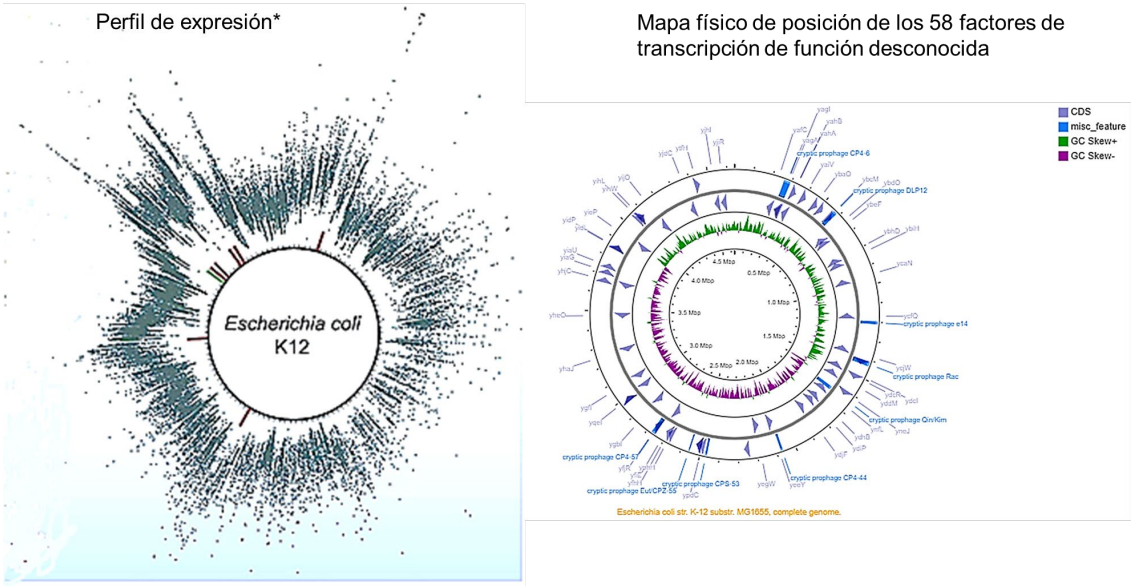


Figura 8. Los niveles de expresión en del genoma de E. coli no son homogéneos, hay zonas menos y más activas. *Tomado y modificado de Scholz *et al*., 2019.

El análisis hecho con la conservación de secuencias mediante árboles filogenético y comparaciones estructurales de los modelos hechos, los resultados sugieren que existe una conservación en estos factores de transcripción no solamente a nivel de la familia de factores transcripcionales que se predice para cada uno de ellos pertenecen, también presentan conservación de otros elementos que los distinguen como son las secuencias sobrerrepresentadas exclusivas para un grupo de estos factores transcripcionales de función desconocida, elementos en el promotor que sugieren ser adquiridos por transferencia horizontal y una clara desviación en el contenido de G+C en la tercera posición de cada codón que los codifica sin variar de manera importante su índice de adaptación de codones (CAI). Estos resultados ayudan a entender el perfil de genes foráneos ya que algunas características incluyen cambios en el uso de codones y el porcentaje de contenido de G+C, así como la presencia de elementos móviles

como son las secuencias de inserción o secuencias repetidas o hot-spots como los sitios donde hay tRNAs (da Silva *et al.*, 2018, Germon *et al.*, 2007)

Una evidencia comúnmente asociada a movimiento de zonas de los genomas son las islas de patogenicidad, definidas como el conjunto de genes necesarios para que un microorganismo bacteriano se comporte como un patógeno (Desvaux *et al.*, 2020), siendo este tipo de análisis los más usados para evaluar la movilidad genómica (Tassinari *et al.*, 2020). Sin embargo, poco se sabe de otros elementos móviles, entre ellos, factores de transcripción, que sean adquiridos de manera independiente o en grupos mediante mecanismos de transferencia horizontal.

Si bien se tiene mucha información de los genomas de bacterias entéricas y se ha podido inferir muchas características de estos, especialmente su diversidad y dinámica (Lukjancenko *et al.*, 2010, Dobrindt *et al.*, 2010). Además, hay muchos elementos que llaman la atención, como es diversidad en el tamaño de los genomas aún de la misma especie, por ejemplo, la cepa BL21 muy usada en experimentos de expresión de proteínas recombinantes es el genoma de E. coli más pequeño encontrado, con tan solo 4.56 Mpb, el número de genes, el tamaño de los genes y el tipo de secuencias reguladoras (Lukjancenko *et al.*, 2010).

El análisis de la micro diversidad en los genomas bacterianos, constituyen elementos para analizar la evolución de genes y genomas, mecanismos de transferencia horizontal y los mecanismos de diversidad tanto local como general de los genomas (Touzain *et al.*, 2010). Un ejemplo muy estudiado son los locus de genes que codifican para tRNAs y RNAs pequeños (Germon *et al.*, 2007, Sridhar y Rafi, 2007), siendo sitios de alta movilidad genómica. Los resultados presentados en este trabajo sugieren que algunos elementos más esenciales potencialmente son también centros de diversidad genómica. El análisis de sintenia, muestra diversidad en las cepas de E. coli, lo que es interesante ya que se esperaría poca variación entre ellas. Incluso muestran re-arreglos genómicos.

El análisis de las secuencias promotoras sugiere que estos genes potencialmente son silenciados en la mayoría de las condiciones mediante H-NS, dada la riqueza de las secuencias en A+T (Lang *et al.*, 2007). Sin embargo, se requiere analizar a más detalle los perfiles de expresión reportados en GenExpDB para evaluar y correlacionar las condiciones experimentales reportadas en todos los experimentos de microarreglos el comportamiento de los 58 factores de transcripción analizados en este trabajo.

Recientemente se ha demostrado que la comparación de varias cepas y especies del género Escherichia usando métodos genéticos permite encontrar nueva esencialidad de genes, es decir, aquellos que en ciertas condiciones de crecimiento son requeridos (Rousset *et al.*, 2021). En este trabajo se ha demostrado que para que un gen sea esencial se requiere de otros elementos como es el efecto epistático de elementos móviles e incluso pueden activar que algunos genes sean esenciales cuando normalmente no lo son. Una posible expansión a este trabajo será evaluar el comportamiento de estos 58 factores transcripcionales en más cepas de E. coli y saber qué efecto tiene la pérdida de algunos de estos en condiciones diversas de crecimiento.

Una de las metas modernas de la biología es la reducción de genomas de organismos modelo, con el fin de tener fábricas de metabolitos de interés biomédico. Reducir el genoma de E. coli ha implicado numerosos retos, el principal es la caída en su capacidad reproductiva, junto con otros problemas de estabilidad genómica (Kurokawa y Ying, 2019). Aun así, el conocimiento de la función de más genes y su procedencia, permitiría coadyuvar con el diseño de células mínimas. Un ejemplo es el factor transcripcional ygbI está ausente en la cepa con genoma sintético mínimo MS56 (genoma de 1.1 Mbp), de manera espontánea, en experimentos de evolución adaptativa en laboratorio o ALE (por sus siglas en inglés), eliminó un fragmento de 21 Kb después de 352 generaciones, lo que le permite recuperar su crecimiento normal debido a la pérdida de rpoS y mutS, en la misma región genómica (Choe etl al., 2019). Por tanto, entre más información se tenga del origen y posible función de genes, se pude diseñar mejores estrategias para generar genomas reducidos.

A manera de conclusión, aun existiendo métodos automatizados para analizar la literatura publicada para obtener datos más robustos sobre los factores de transcripción (Méndez-Cruz *et al.*, 2020), se requieren de métodos cuidadosos y vigilados por humanos para evaluar nuevas propiedades en las secuencias de proteínas aún ya reportadas y anotadas. Este trabajo es un ejemplo de cómo se pueden encontrar nuevos aspectos en secuencias ya reportadas y que presentan características relevantes y se tiene suficiente evidencia para poder decir que la mayoría de los 58 factores transcripcionales tienen un origen reciente en el genoma de E. coli y potencialmente se han ido adaptando a ser funcionales en este organismo.

Finalmente, es una posibilidad que algunos de estos factores transcripcionales tengan una función que no ha sido caracterizada ya que existen ejemplos de factores transcripcionales adquiridos de manera horizontal y que tienen una función especial, tal es el caso del regulador GmrA en E. coli O157:H7, una cepa patógena que contiene este regulador y controla la expresión de los genes flagelares (Yang *et al.*, 2018).

Referencias

1. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K.A., Tomita, M., Wanner, B.L., Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2:2006.0008. doi: 10.1038/msb4100050.

2. Blattner, F.R., Plunkett, G. 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., *et al.* (1997) The complete genome sequence of Escherichia coli K-12. Science, 277:1453-74. 1.

3. Boël, G., Letso, R., Neely, H., Price, W.N., Wong, K.H., Su, M., Luff, J., Valecha, M., Everett, J.K., Acton, T.B., Xiao, R., Montelione, G.T., Aalberts, D.P., Hunt, J.F. (2016). Codon influence on protein expression in E. coli correlates with mRNA levels. Nature. 21;529(7586):358-363. doi: 10.1038/nature16509.

4. Bykov, A., Glazunova, O., Alikina, O., Sukharicheva, N., Masulis, I., Shavkunov, K., Ozoline, O. (2020). Excessive Promoters as Silencers of Genes Horizontally Acquired by Escherichia coli. Front Mol Biosci. 26;7:28. doi: 10.3389/fmolb.2020.00028.

5. Choe, D., Lee, J.H., Yoo, M., Hwang, S., Sung, B.H., Cho, S., Palsson, B., Kim, S.C., Cho, B.K. (2019). Adaptive laboratory evolution of a genome-reduced Escherichia coli. Nat Commun. 2019 Feb 25;10(1):935. doi: 10.1038/s41467-019-08888-6.

6. da Silva Filho, A.C., Raittz, R.T., Guizelini, D., De Pierri, C.R., Augusto, D.W., Dos Santos-Weiss, I.C.R., Marchaukoski, J.N. (2018). Comparative Analysis of Genomic Island Prediction Tools. Front Genet. 12;9:619. doi: 10.3389/fgene.2018.00619.

7. Daubin, V., Ochman, H. (2004). Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. Genome Res. 14 1036-1042. 10.1101/gr.2231904.

8. de Jong, A., Pietersma, H., Cordes, M., Kuipers, O.P., Kok, J. (2012). PePPER: a webserver for prediction of prokaryote promoter elements and regulons. BMC Genomics. 2;13:299. doi: 10.1186/1471-2164-13-299.

9. Desvaux, M., Dalmasso, G., Beyrouthy, R., Barnich, N., Delmas, J., Bonnet, R. (2020). Pathogenicity Factors of Genomic Islands in Intestinal and Extraintestinal Escherichia coli. Front Microbiol. 2020 Sep 25;11:2065. doi: 10.3389/fmicb.2020.02065.

10. Dobrindt, U., Chowdary, M.G., Krumbholz, G., Hacker, J. (2010). Genome dynamics and its impact on evolution of Escherichia coli. Med Microbiol Immunol. 199(3):145-54. doi: 10.1007/s00430-010-0161-2.

11. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 32(5):1792-1797

12. Germon, P., Roche, D., Melo, S., Mignon-Grasteau, S., Dobrindt, U., Hacker, J., Schouler, C., Moulin-Schouleur, M. (2007). tDNA locus polymorphism and ecto-chromosomal DNA insertion hot-spots are related to the phylogenetic group of Escherichia coli strains. Microbiology (Reading). 53(Pt 3):826-837. doi: 10.1099/mic.0.2006/001958-0.

13. Gilson, E., Clément, J., Brutlag, D. and Hofnung, M. (1984). A family of dispersed repetitive extragenic palindromic DNA sequences in E. coli. The EMBO Journal, 3: 1417-1421. https://doi.org/10.1002/j.1460-2075.1984.tb01986.x.

14. Grant, J.R., Stothard, P. (2008) The CGView Server: a comparative genomics tool for circular genomes. Nucleic Acids Res. 1;36(Web Server issue):W181-4. doi: 10.1093/nar/gkn179.

15. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., Noble, W.S. (2007). Quantifying similarity between motifs. Genome Biology, 8(2):R24. doi: 10.1186/gb-2007-8-2-r24.

16. Huang, Q., Cheng, X., Cheung, M. K., Kiselev, S. S., Ozoline, O. N., Kwan, H. S. (2012). High density transcriptional initiation signals underline genomic islands in bacteria. PLoS One 7:e33759.

17. Keseler, I.M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martinez, C., Caspi, R., Fulcher, C., Gama-Castro, S., Kothari, A., Krummenacker, M., Latendresse, M., Muñiz-Rascado, L., Ong, Q., Paley, S., Peralta-Gil, M., Subhraveti, P., Velazquez-Ramirez, D.A., Weaver, D., Collado-Vides, J., Paulsen, I., and Karp, P.D. (2017). The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. Nucleic Acids Research 45:D543-550.

18. Kurokawa, M., Ying, B.W. (2019). Experimental Challenges for Reduced Genomes: The Cell Model Escherichia coli. Microorganisms. 18;8(1):3. doi: 10.3390/microorganisms8010003.

19. Lang, B., Blot, N., Bouffartigues, E., Buckle, M., Geertz, M., Gualerzi, C.O., Mavathur, R., Muskhelishvili, G., Pon, C.L., Rimsky, S., Stella, S., Babu, M.M., Travers, A. (2007). High-affinity DNA binding sites for H-NS provide a molecular basis for selective silencing within proteobacterial genomes. Nucleic Acids Res. 35(18):6330-7. doi: 10.1093/nar/gkm712.

20. Lavner, Y., Kotlar, D. (2005). Codon bias as a factor in regulating expression via translation rate in the human genome. Gene. 345: 127-138.

21. Lukjancenko, O., Wassenaar, T.M., Ussery, D.W. (2010). Comparison of 61 sequenced Escherichia coli genomes. Microb Ecol. 60(4):708-20. doi: 10.1007/s00248-010-9717-3.

22. Méndez-Cruz, C.F., Blanchet, A., Godínez, A., Arroyo-Fernández, I., Gama-Castro, S., Martínez-Luna, S.B., González-Colín, C., Collado-Vides, J. (2020). Knowledge extraction for assisted curation of summaries of bacterial transcription factor properties. Database (Oxford). 11;2020:baaa109. doi: 10.1093/database/baaa109.

23. Purtov, Y.A., Glazunova, O.A., Antipov, S.S., Pokusaeva, V.O., Fesenko, E.E., Preobrazhenskaya, E.V., Shavkunov, K.S., Tutukina, M.N., Lukyanov, V.I., Ozoline, O.N. (2014). Promoter islands as a platform for interaction with nucleoid proteins and transcription factors. J Bioinform Comput Biol. 12(2):1441006. doi: 10.1142/S0219720014410066.

24. Riley, M., Labedan, B. (1997). Protein evolution viewed through Escherichia coli protein sequences: introducing the notion of a structural segment of homology, the module. J Mol Biol. 23;268(5):857-68. doi: 10.1006/jmbi.1997.1003.

25. Rousset, F., Cabezas-Caballero, J., Piastra-Facon, F., Fernández-Rodríguez, J., Clermont, O., Denamur, E., Rocha, E.P.C., Bikard, D. (2021). The impact of genetic diversity on gene essentiality within the Escherichia coli species. Nat Microbiol. 6(3):301-312. doi: 10.1038/s41564-020-00839-y.

26. Scholz, S.A., Diao, R., Wolfe, M.B., Fivenson, E.M., Lin, X.N., Freddolino, P.L. (2019). High-Resolution Mapping of the Escherichia coli Chromosome Reveals Positions of High and Low Transcription. Cell Syst. 27;8(3):212-225.e9. doi: 10.1016/j.cels.2019.02.004.

27. Sridhar J, Rafi ZA. (2007). Identification of novel genomic islands associated with small RNAs. In Silico Biol. 7(6):601-11.

28. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., Jensen, L.J., Mering, C.V. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 8;47(D1):D607-D613. doi: 10.1093/nar/gky1131

29. Tassinari, E., Bawn, M., Thilliez, G., Charity, O., Acton, L., Kirkwood, M., Petrovska, L., Dallman, T., Burgess, C.M., Hall, N., Duffy, G., Kingsley, R.A. (2020). Whole-genome epidemiology links phage-mediated acquisition of a virulence gene to the clonal expansion of a pandemic Salmonella enterica serovar Typhimurium clone. Microb Genom. 6(11):mgen000456. doi: 10.1099/mgen.0.000456.

30. Touzain, F., Denamur, E., Médigue, C., Barbe, V., El Karoui, M., Petit, M.A. (2010). Small variable segments constitute a major type of diversity of bacterial genomes at the species level. Genome Biol. 11(4):R45. doi: 10.1186/gb-2010-11-4-r45.

31. Wang, S., Ma, J., Peng, J., Xu, J. (2013). Protein structure alignment beyond spatial proximity. Scientific Reports. 3:1448. doi: 10.1038/srep01448.

32. Wang, S., Peng, J., Xu, J. (2011) Alignment of distantly related protein structures: algorithm, bound and implications to homology modeling. Bioinformatics. 27(18):2537-45. doi: 10.1093/bioinformatics/btr432.

33. Yamamoto, K., Hirao, K., Oshima, T., Aiba, H., Utsumi, R., Ishihama, A. (2005) Functional characterization in vitro of all two-component signal transduction systems from Escherichia coli. J Biol Chem. 14;280(2):1448-56. doi: 10.1074/jbc.M410104200.

Volumen 10
XXVI Verano de la Ciencia
ISSN 2395-9797
www.jovenesenlaciencia.ugto.mx

34. Yamamoto, N., Nakahigashi, K., Nakamichi, T., Yoshino, M., Takai, Y., Touda, Y., Furubayashi, A., Kinjyo, S., Dose, H., Hasegawa, M., Datsenko, K.A., Nakayashiki, T., Tomita, M., Wanner, B.L., Mori, H. (2009). Update on the Keio collection of Escherichia coli single-gene deletion mutants. Mol Syst Biol. 5:335. doi: 10.1038/msb.2009.92.

35. Yang, B., Wang, S., Huang, J., Yin, Z., Jiang, L., Hou, W., Li, X., Feng, L. (2018). Transcriptional Activator GmrA, Encoded in Genomic Island OI-29, Controls the Motility of Enterohemorrhagic Escherichia coli O157:H7. Front Microbiol. 22;9:338. doi: 10.3389/fmicb.2018.00338.