2022

Evaluación de la posible existencia biológica de proteínas a partir de secuencias de ARNs generados por modelamiento computacional pseudoaleatorio

Joan Sebastián Gutiérrez Sánchez
Universidad de La Salle, Bogotá, jgutierrez19@unisalle.edu.co

Andrés Reinaldo Chacón Prada
Universidad de La Salle, Bogotá, achacon30@unisalle.edu.co

Evaluación de la posible existencia biológica de proteínas a partir de secuencias de ARNs generados por modelamiento computacional pseudoaleatorio.

Joan Sebastián Gutiérrez Sánchez

Andrés Reinaldo Chacón Prada

Programa de Biología, Departamento de Ciencias Básicas, Universidad de La Salle

Trabajo de grado para optar al título de biólogo.

Diana Carolina Ochoa Cabezas

Ph.D Abelino Vargas Jimenes

25 de mayo de 2022

Notas de Autor

Agradecemos a todas las personas que hicieron posible el desarrollo de esta investigación especialmente a nuestros tutores por brindarnos parte de su valioso tiempo.

Cualquier inquietud respecto a este documento de investigación debe ser remitida al programa de biología de la Universidad de La Salle, Bogotá, Colombia.

E-mails: jgutierrez19@unisalle.edu.co & achacon30@unisalle.edu.co

Tabla de Contenidos

Lista de Figuras

Lista de Tablas

Resumen

Las proteínas son biomoléculas fundamentales para el funcionamiento de los sistemas biológicos, por lo que entender como surgen y evolucionan es de gran interés teórico. Algunos autores consideran que el origen de las proteínas se dio por el ordenamiento aleatorio de secuencias polipeptídicas; por este motivo el objetivo de este trabajo es inferir si el proceso de creación de secuencias de ARN mensajeros es de carácter estocástico, mediante el diseño y programación de un código computacional en Python que genera secuencias de ARN de manera pseudoaleatoria; posteriormente, se tradujeron las secuencias de ARN obtenidas a aminoácidos para poder realizar un BLAST en busca de homologías; con estos datos se realizó un análisis probabilístico con los porcentajes de identidad arrojados por el programa en una muestra aleatoria de 100 secuencias con un porcentaje de identidad mayor al 30%. Este análisis mostró que más de la mitad de las secuencias generadas pseudoaleatoriamente presentaban homologías con proteínas encontradas en la naturaleza, y a través del análisis probabilístico se pudo inferir que en cualquier muestra aleatoria de 100 secuencias se tiene una probabilidad del 0.45 de que el promedio del porcentaje de identidad sea mayor al 30%, demostrando así que es muy probable encontrar homologías en cualquier muestra aleatoria de secuencias formadas estocásticamente, por lo tanto, muchas de estas pueden generar plegamientos estructurales y desempeñar funciones específicas, de esta manera aportando evidencia a favor de la posibilidad del origen estocástico de las proteínas.

Palabras clave: código pseudoaleatorio, homología, secuencia aleatoria, porcentaje de identidad, proteínas.

Evaluación de la posible existencia biológica de proteínas a partir de secuencias de ARNs generados por modelamiento computacional pseudoaleatorio.

Introducción

Las proteínas son biomoléculas de gran complejidad química que permiten el desarrollo de la estructura y la funcionalidad de los sistemas biológicos (Lodish et al., 2008); es por esto, que entender la naturaleza del proceso mediante el cual se seleccionan y cambian los componentes plasmados en el código genético que dan paso a la formación de proteínas, es de gran interés teórico para la biología. Esto se debe a que dicho proceso está estrechamente relacionado con el origen y la evolución de estas biomoléculas presentes en todas las formas de vida (Delport et al., 2010).

Una característica de las proteínas desde el punto de vista estructural, funcional y evolutivo es su naturaleza modular. Se considera que los módulos básicos que conforman a las proteínas son los dominios proteínicos que, por un lado, presentan plegamientos tridimensionales independientes y, por otro lado, son fragmentos conservados evolutivamente (Moore et al., 2008). La composición modular de las proteínas es considerado como un aspecto clave en la evolución de proteínas preexistentes y en el surgimiento de nuevas cadenas polipeptídicas. Gracias al reordenamiento de un repertorio finito de módulos se produce la gran variedad de funcionalidades observadas en las proteínas, por lo que, la evidencia sugiere que en los organismos vivos existe una preferencia por la reutilización de estos módulos previamente presentes, en vez del uso de módulos que surgen *de novo* (Bornberg-Bauer et al., 2005). Un posible mecanismo mediante el cual se producen las proteínas con múltiples dominios es a través de la adición o eliminación de dominios en los extremos de la secuencia de aminoácidos, esta adición o eliminación se da como consecuencia de una serie de eventos a nivel genético, como la

duplicación genética o la fusión y fisión de genes. Además, las mutaciones puntuales que producen codones de parada prematuros en las secuencias generan que se pierda el resto terminal de las mismas ocasionando la posible reorganización modular de las proteínas (Pasek et al., 2006; Weiner et al., 2006).

Una particularidad sobre los módulos encontrados en las proteínas es la caracterización de dominios huérfanos, los cuales no tienen homólogos conocidos. Los dominios huérfanos tienen un alto porcentaje de regiones desordenadas, es decir, que no se pliegan en estructuras secundarias; también poseen un mayor porcentaje de regiones de baja complejidad; en donde la baja complejidad hace referencia a regiones con distribuciones no aleatorias de aminoácidos, en comparación con dominios con homólogos conocidos (Brown et al., 2002; Ekman et al., 2005).

Como ya se ha mencionado, las proteínas están constituidas por un repertorio finito de módulos, en su mayoría previamente existentes, por lo que se plantea la incógnita de ¿cómo surgieron en un principio los módulos de los cuales hacen uso las proteínas en la actualidad? (Trifonov y Frenkel, 2009). Es posible que la organización de la información que daría paso a la conformación de los módulos proteicos se diera de manera aleatoria y posteriormente se seleccionarían gracias a las ventajas otorgadas a los organismos portadores, por lo que lograron permanecer hasta la actualidad. Algunas teorías sobre el origen aleatorio de los módulos proteicos están basadas en la naturaleza estocástica de las células, debido al carácter discreto y al poco número de moléculas que intervienen en los procesos intracelulares, lo que genera fluctuaciones que son intrínsecas a los comportamientos biológicos (Chess, 2013; Jensen et al., 2020; Saiz y Vilar, 2006). Por ejemplo, en la generación de proteínas, aparecen variaciones aleatorias en el número de moléculas y por ende en la velocidad de reacción en cada una de las etapas de transcripción y traducción. Es por esto por lo que la aleatoriedad no se puede atribuir a

limitaciones en la descripción, si no a una propiedad fundamental de los sistemas biológicos (Dessalles, 2017).

Para entender el cambio de las bases nitrogenadas y por ende el cambio de los aminoácidos que componen las proteínas, se han propuesto diversos mecanismos que conducirían la evolución a nivel molecular. Algunos autores proponen que el cambio de los aminoácidos se da por selección positiva, y por ende sería necesario desarrollar una teoría de selección natural a nivel molecular (Kelley y Swanson, 2008), pero se han encontrado limitaciones en los métodos desarrollados para apoyar las observaciones que justifican la selección natural a nivel molecular (Nei et al., 2010), por otra parte, la teoría neutral de la evolución molecular afirma que los cambios evolutivos de los biopolímeros que ocurren en los sistemas biológicos se dan en su mayoría por la fijación aleatoria de mutantes que son selectivamente neutrales; esta alternativa pone énfasis en el rol que tiene la presión mutacional y la deriva genética aleatoria. Aunque las substituciones de los elementos del ADN y por ende de las proteínas ocurren de manera aleatoria, esto no malogra la información genética, por lo que la ocurrencia de la neutralidad no ocurre solo en secciones del genoma que no codifican para proteínas, sino también en sitios importantes para la funcionalidad proteínica (Kimura, 1991).

Además de las teorías de evolución molecular se han propuesto diferentes modelos teóricos que permiten estudiar la dinámica evolutiva de las proteínas, uno de estos modelos son los paisajes moleculares de fitness que están conformados por un espacio de secuencias y por una función que determina el fitness de cada secuencia que se encuentra en este espacio. Cuando el cambio de una secuencia a otra está dado por mutaciones completamente estocásticas se puede considerar a la evolución molecular como una caminata aleatoria en el espacio de secuencias, el espacio de secuencias es un objeto matemático discreto con N dimensiones, donde N

corresponde al número de sitios (monómeros) variables del biopolímero, es decir, la longitud de la secuencia; a su vez, cada una de las N dimensiones posee m puntos que corresponden al tamaño del alfabeto que conforma las secuencias. Por ejemplo, el ADN posee un alfabeto conformado por 4 símbolos (a,g,c,t). Estas características hacen que el espacio de secuencias en general sea inmensamente grande ya que posee un tamaño de m^N dificultando su estudio de manera *in vitro* o *in vivo*, por otra parte, el fitness es una variable continua que en su mayoría se define a través de observaciones experimentales (Blanco et al., 2019; Fragata et al., 2019).

Algunos autores consideran que, en el origen de los biopolímeros, las proteínas primordiales capaces de desarrollar funcionalidad se pudieron haber generado a partir de secuencias aleatorias de aminoácidos, donde estas secuencias *de novo* carecían de historia evolutiva, pero a su vez se cree que la ocurrencia de proteínas con funciones específicas en el espacio de secuencias es muy rara (Tong et al., 2021). Keefe y Szostak (2001) realizaron un experimento en donde exploraron a través de la selección *in vitro* de un espacio de secuencias formado por una librería aleatorizada de tamaño 6×10^{12}, en este estudio encontraron que con una frecuencia de $\approx 1\ en\ 10^{11}$ aparecieron proteínas con una funcionalidad especifica, en este caso la capacidad de unirse a la molécula de ATP; sugiriendo así que las proteínas funcionales son lo suficientemente comunes para ser encontradas a través de procesos completamente estocásticos, lo que plantea el escenario de que de esta manera pudo haber ocurrido en las primeras formas de vida.

Por otra parte, se ha mostrado que secuencias de aminoácidos formadas aleatoriamente expresadas en células de *Escherichia coli* presentan plegamientos con propiedades similares a proteínas nativas, además se evidenció que estas secuencias aleatorias poseían gran resistencia a la denaturación, mostrando así gran estabilidad contra diferentes perturbaciones ambientales y en

algunos casos mostrando algún tipo de funcionalidad (Davidson y Sauer, 1994). Adicionalmente a estos estudios, se ha mostrado en varias ocasiones que proteínas *de novo* formadas por secuencias aleatorias sin historia evolutiva presentan características estructurales, funcionales y fisicoquímicas similares a las propiedades observadas en las proteínas actuales (Tong et al., 2021); además, se tiene conocimiento que uno de los mecanismos mediante el cual surgen nuevos genes codificadores de proteínas funcionales es por la combinación aleatoria de fragmentos dispersos del genoma (Bornberg-Bauer et al., 2010).

Es en este contexto, la exploración de espacios de secuencias conformados por cadenas aleatorias de aminoácidos en busca de una fracción de ellas que posea funcionalidad biológica, permitiría deducir si las proteínas que observamos en la actualidad pudieron haber surgido por la formación estocástica de polipéptidos a partir de secuencias aleatorias de ARN, lo que también ayudaría a entender como surgen nuevas proteínas, ya que sabemos que el número de secuencias de aminoácidos diferentes es astronómicamente grande y que la naturaleza solo puede explorar una fracción pequeña de todas las secuencias posibles, es por lo anterior, que esta investigación busca responder la siguiente pregunta: ¿Es posible que las proteínas se hayan originado por la combinación aleatoria de aminoácidos? Mediante el uso del programa computacional desarrollado en este estudio se podría ayudar a entender como surgen y evolucionan los ARN mensajeros que posteriormente se traducen en proteínas, ya que a través del código generamos una marcha aleatoria en un espacio de secuencias extremadamente grande que sería imposible de estudiar en condiciones de laboratorio, esto permitió realizar una exploración de secuencias posibles para posteriormente evaluar si las secuencias obtenidas y las proteínas generadas a partir de estas poseen zonas homologas con secuencias encontradas en la naturaleza, demostrando así,

que proteínas generadas estocásticamente presentan funcionalidad biológica lo que apoyaría la hipótesis del origen estocástico de estas biomoléculas.

Objetivos

Objetivo General

Inferir si el proceso creación de secuencias de ARN mensajeros es de carácter estocástico.

Objetivos Específicos

- Diseñar un código pseudoaleatorio para generar secuencias de ARN.
- Establecer si las secuencias de ARNs generadas producen polipéptidos que tengan homologías con proteínas reportadas en la naturaleza
- Evaluar si las secuencias de proteínas que se conocen en la actualidad pudieron tener un origen estocástico.

Materiales y Métodos

Diseño Experimental

El experimento se realizó 100% *in silico* usando un ordenador portátil Lenovo que cuenta con un procesador AMD FX-7500 Radeon R7, 10 Compute Cores 4C+6G a 2.10 GHz y una memora RAM de 4 Gb. Como el código desarrollado es un loop infinito donde no se programaron criterios de parada se tuvieron en cuenta solo las primeras 150 bases nitrogenadas que corresponden a 50 aminoácidos, ya que se conoce que la insulina que es una proteína relativamente pequeña cuenta con 51 aminoácidos, lo cual, indica que con este tamaño mínimo ya se podría obtener una secuencia de aminoácidos que genere una proteína funcional con la ventaja de facilitar su obtención a través del código pseudoaleatorio generado.

Procedimientos por Objetivos Específicos

Primer Objetivo Específico. Por medio de Python (Van Rossum y Drake, 2011) se importó la librería math que permitió computar operadores matemáticos booleanos y la librería random que permitió generar números pseudoaleatorios a partir de una distribución uniforme discreta; después se realizó la definición de igualdad para las variables de la siguiente manera: a=1; u=2; c=3; g=4, donde **a** es adenina; **u** es uracilo; **c** es citocina y **g** es guanina. Posteriormente, se definió que aug=124 como premisa de igualdad verdadera en concordancia con la definición de variables realizada anteriormente, esto permitió realizar un while loop infinito ya que se programó que mientras 1 sea diferente a la condición de igualdad se generen números entre 1 y 4 de manera pseudoaleatoria, así se obtuvo una secuencia infinita pseudoaleatoria de números $x_1x_2x_3x_4\ldots\ tal\ que\ x_i \in$ [1,2,3,4]. Se seleccionaron únicamente las primeras 150 posiciones de cada secuencia infinita. Este código se corrió 100 veces y se registró una secuencia por cada una, con las 100 secuencias obtenidas, se usó la herramienta EMBOSS Transeq (Rice *et al.,* 2000) del EMBL-EBI para realizar la traducción de la secuencia de ARN a una secuencia de aminoácidos. Este procedimiento se realizó con marco de lectura 1, código genético bacteriano y el resto de los parámetros de manera predeterminada.

Segundo Objetivo Específico. Para evaluar la probabilidad de existencia biológica de nuestras secuencias de aminoácidos se realizó un BLAST (The Uniprot Consortium, 2021) para homología de proteínas, esta herramienta permitió encontrar regiones idénticas entre nuestras secuencias proteicas y proteínas almacenadas en las bases de datos arrojándonos información sobre la estructura y la función de nuestras proteínas. El porcentaje obtenido ayudó a identificar qué cantidad de la secuencia y en donde presenta homología y así se evaluó la presencia de motivos y dominios con características ya reportadas en secuencias descritas para algún

organismo. Para realizar el BLAST+ se ingresó a la página web de uniprot y se seleccionó el apartado de BLAST, una vez allí se escogió la opción de proteomas de referencia uniprotKB más Swiss-prot como base de datos para el BLAST y se dejó al resto parámetros de manera predeterminada.

Tercer Objetivo Específico. Una vez realizado el BLAST (The Uniprot Consortium, 2021) de las 100 secuencias se calculó el promedio del porcentaje de identidad obtenido y mediante el teorema del límite central se estandarizó la media y la desviación estándar, posteriormente se realizó el cálculo para determinar cuál es la probabilidad de que el promedio de identidad de 100 secuencias sea mayor al 30%, ya que este es un criterio para determinar la homología (Gómez et al., 2011), así pudimos extrapolar el promedio de identidad obtenido en nuestro grupo de secuencias a cualquier muestra aleatoria de 100 secuencias. La fórmula usada para calcular la probabilidad aproximada fue la siguiente:

Gracias al teorema del límite central, sabemos que $z_{n=(\bar{X}-\mu)/[(S)/(\sqrt{n})]}$ cuando $n \to \infty$ posee una distribución de probabilidad normal estándar, clásicamente se considera que a partir de un $n > 30$ se obtiene una aproximación lo suficientemente buena, por lo que un tamaño muestral de 100 sería adecuado para aplicar el teorema del límite central. Para calcular la probabilidad aproximada donde el promedio del porcentaje de identidad de una muestra de 100 secuencias sea mayor a 0.30 se usa la tabla de la distribución normal estándar de la siguiente manera:

$$P\{\bar{X} > 0.30\} = P\left\{Z > \frac{0.30-\bar{X}}{\frac{S}{\sqrt{n}}}\right\} \quad (1)$$

Para calcular la media muestral se usó la siguiente formula:

$$\bar{X} = \frac{1}{n} \times \sum_{i=1}^{n} X_i \quad (2)$$

Por otra parte, para calcular la desviación estándar se usó la siguiente formula:

$$S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}} \quad (3)$$

Análisis de Datos

El análisis de identidad se realizó usando la herramienta bioinformática del BLAST para proteínas, con los resultados arrojados por el BLAST se realizó una base de datos en Excel que contenía la información de las secuencias de ARNm y de aminoácidos, además de la proteína, el organismo y el porcentaje de identidad correspondiente (apéndice A y B). Usando esta base de datos se extrajo información sobre el número de codones de parada presentes en las secuencias, el tipo de proteína con el que se encontraron identidades y como se distribuían los porcentajes de identidad obtenidos; esta información se graficó usando el software estadístico R y Excel, por otra parte, para analizar los porcentajes de identidad obtenidos se realizó una base de datos que contenía los valores de identidad y los valores necesarios para calcular la desviación estándar de la muestra (apéndice C), una vez obtenidos la media y la desviación estándar se usó la ecuación 1 para calcular la probabilidad deseada usando el teorema del límite central y la tabla de distribución normal estándar y con esta probabilidad se pudo extrapolar los resultados estadísticos obtenidos en nuestra muestra a cualquier muestra aleatoria de 100 secuencias.

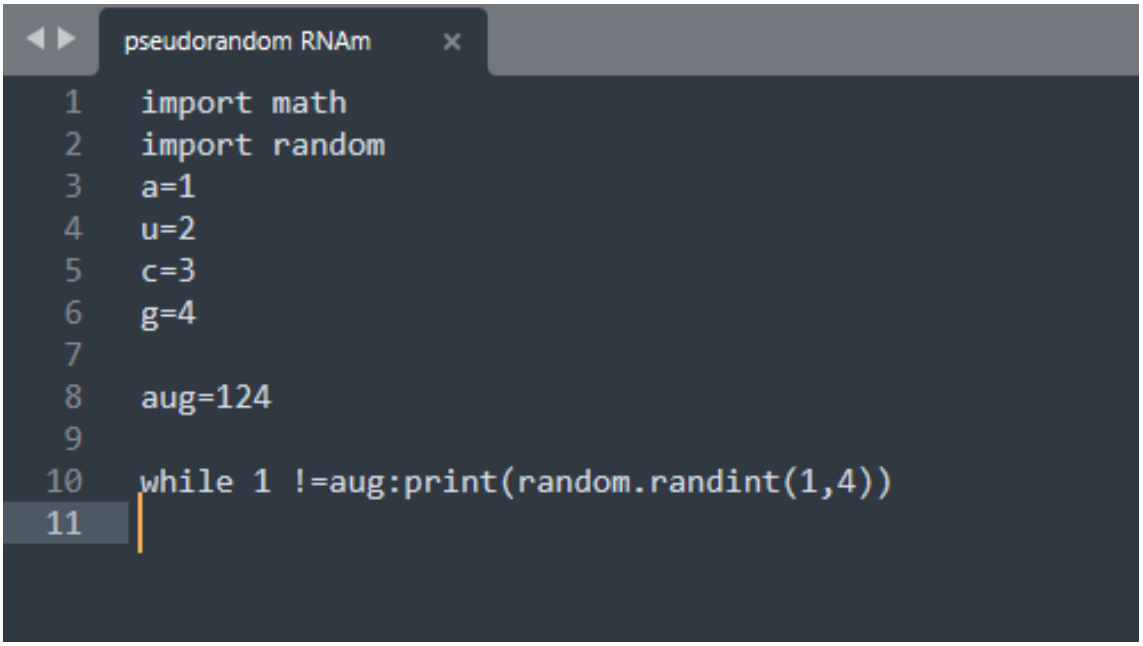
Resultados

Programa Computacional Pseudoaleatorio

Se logró generar un código computacional pseudoaleatorio a través de un while loop que producía secuencias infinitas conformadas por los números 1,2,3,4 (Figura 1)

Figura 1

Código Pseudoaleatorio en Python



Nota: en esta figura se describe en detalle el código computacional elaborado para formar secuencias de ARN pseudoaleatorias.

Al correr el código 100 veces distintas se obtuvieron 100 secuencias diferentes en donde por cada vez que se corría el programa se seleccionaban los primeros 150 elementos de la secuencia infinita, al fijar el tamaño de las secuencias en 150 se estableció un espacio de secuencias de tamaño 4^{150}. El programa realiza una caminata aleatoria en este espacio de secuencias con una probabilidad igual entre cada paso de la caminata, ya que cada uno de los

cuatro componentes de las secuencias tiene la misma probabilidad de 0.25 de ocurrencia, por lo que, el resultado que se obtiene al correr el código 100 veces es un muestreo aleatorio de 100 secuencias del espacio de secuencias.

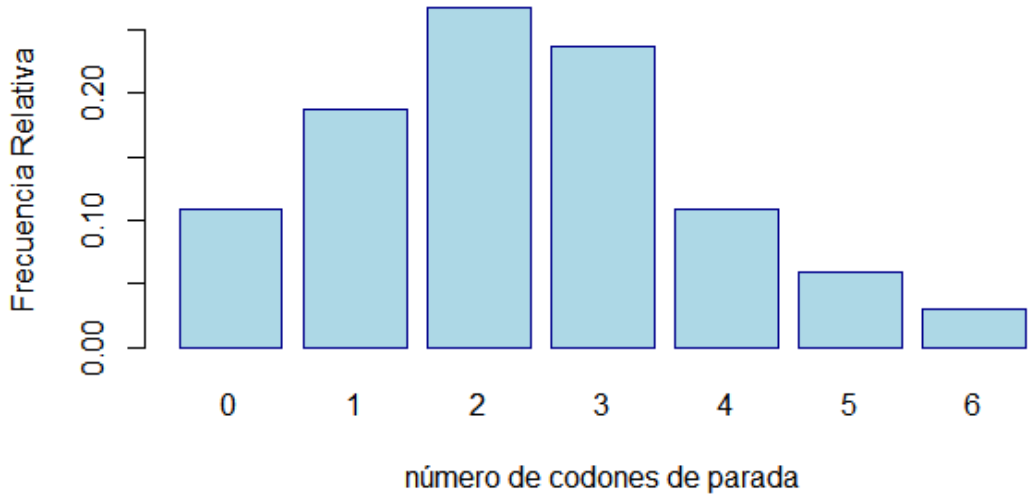
Las 100 secuencias obtenidas presentan una proporción similar de a,u,c,g (Apéndice A), esto se da debido a la ley de los grandes números, ya que en una muestra aleatoria se observa que la frecuencia de ocurrencia de cada posible evento tiende a ser igual al valor esperado de cada evento del espacio muestral a medida que el tamaño de la muestra tienda a infinito, es por esto por lo que se espera que la frecuencia de ocurrencia de a,u,c,g en las secuencias obtenidas sea ≈ 0.25.

Secuencias de Aminoácidos

Una vez realizada la traducción de las secuencias de ARNm a secuencias de aminoácidos, se encontró que la gran mayoría de secuencias presentaban codones de parada, donde la aparición de dos codones fue el evento más frecuente, seguida de tres codones y un codón consecutivamente, el número de secuencias que no presentaron codones de parada es igual al número de secuencias que presentaron 4 codones de parada, es decir 11; las secuencias que tiene 5 y 6 codones de parada son las menos frecuentes. En la muestra aleatoria de 100 secuencias de aminoácidos con 50 posiciones no se encontraron secuencias que tuvieran más de 6 codones de parada (Figura 2)

Figura 2

Frecuencia de Aparición de Codones de Parada en Secuencias Aleatorias de 50 Aminoácidos



Nota: en esta figura se muestra la frecuencia de la cantidad de codones de parada encontrados en las secuencias de aminoácidos obtenidas usando el código pseudoaleatorio.

Identidad de las proteínas obtenidas con respecto a proteínas encontradas en la naturaleza

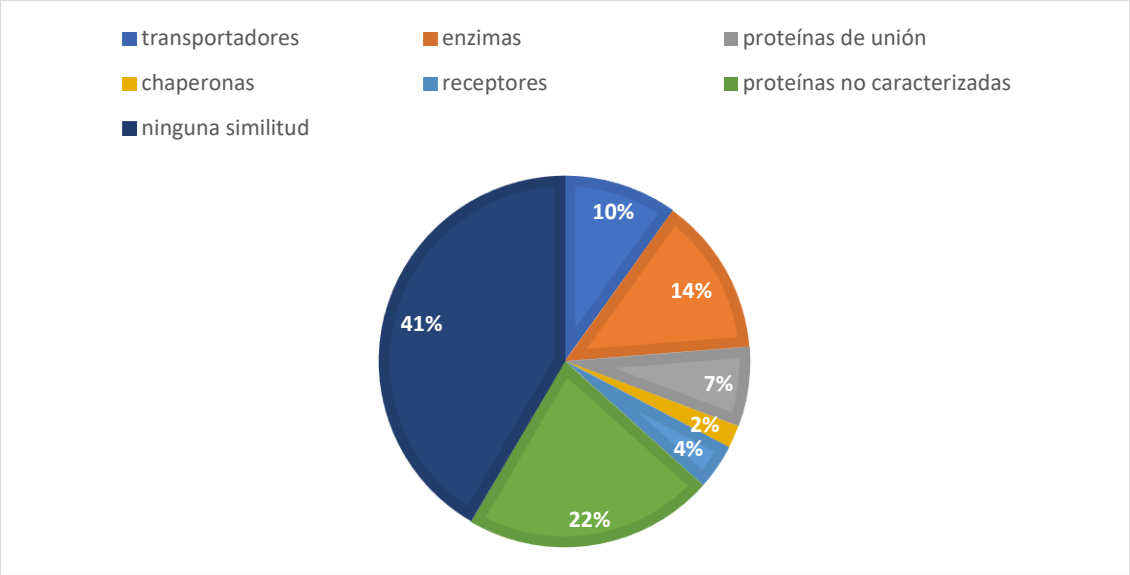
Se encontró que el 59% de las secuencias presentaron identidad, dentro de este grupo se observó la predominancia de secuencias con identidad a proteínas no caracterizadas, seguidas por secuencias con identidad a enzimas, transportadores y posteriormente, en menor medida, secuencias con identidad a proteínas de unión, receptores y chaperonas respectivamente (Figura 3).

Dentro del grupo mencionado anteriormente se encontró identidad con proteínas con funciones claves para los sistemas biológicos, por ejemplo, en cuanto a enzimas se encontró identidad con subunidades de la ADN polimerasa y de la ARN polimerasa, también con dominios con propiedades de cinasa e hidrolasa entre otros y otras involucradas en el metabolismo como fructosa-difosfato aldosa; entre los transportadores se encontró identidad con subunidades del citocromo, proteína fundamental en el proceso de respiración celular; también se

halló identidad con transportadores de iones como el sodio, fundamentales para la estabilidad del potencial de acción celular; dentro de las proteínas de unión se encontraron similitudes con dominios con capacidad de unirse al ADN cumpliendo funciones de factores de transcripción como los dedos de cinc, también proteínas importantes para la biología bacteriana como el domino ompA; por el lado de los receptores se obtuvo identidad con receptores tipo Toll, que son fundamentales para el funcionamiento del sistema inmune innato, por otra parte, se encontró identidad con proteínas que contenían el dominio DNAJ_C cuya funcionalidad se ha relacionado con el de una chaperona (Apéndice B).

Figura 3

Tipos de Proteínas con las que se Encontró Identidad

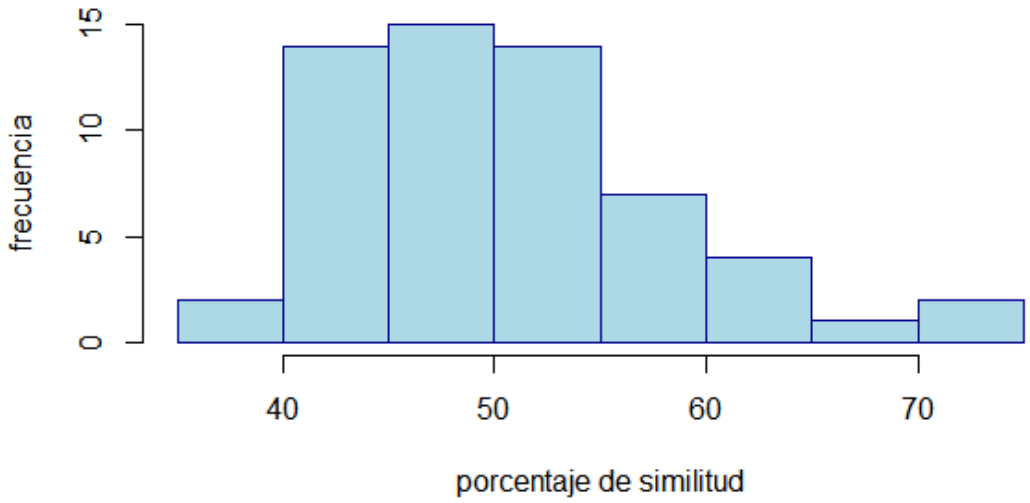


Nota: La figura muestra los tipos de proteínas clasificadas según su función molecular con las que se encontró identidad a través del BLAST. El porcentaje indica la frecuencia relativa de aparición de identidad de cada tipo de proteína en la muestra de 100 secuencias.

En cuanto a los porcentajes de identidad, se encontró que todas las secuencias con las que se obtuvieron resultados en el BLAST superaron el 30%, ya que se distribuyeron entre el 35 y el 75 por ciento, además se encontró que la mayoría de las secuencias que mostraron identidad tuvieron un porcentaje entre 40% y 60% (Figura 4).

Figura 4

Porcentaje de Identidad de las Secuencias Formadas Usando el Código Pseudoaleatorio



Nota: La figura muestra la frecuencia absoluta de los porcentajes de identidad obtenidos.

Evaluación Probabilística de la Posibilidad del Origen Estocástico de las Proteínas

Usando la ecuación 2 se obtuvo que la media del porcentaje de identidad de nuestra muestra aleatoria de 100 secuencias es de 0.296, esta media se usó para calcular los datos requeridos para hallar la desviación estándar (columna 4 del apéndice C). Usando la ecuación 3 se encontró que la desviación estándar de nuestra muestra aleatoria es de 0.261. Con la media y

la desviación estándar se hicieron los siguientes dos cálculos de probabilidad usando la ecuación 1:

1) Probabilidad de que la media de una muestra aleatoria de 100 secuencias sea mayor al 30%.

$$P\{\bar{X} > 0.30\} = P\left\{Z > \frac{0.30 - 0.29}{\frac{0.26}{\sqrt{100}}}\right\}$$

$$= P\{Z > 0.138\}$$

$$= 1 - P\{Z \leq 0.138\}$$

$$= 1 - 0.55$$

$$= 0.45$$

Se encontró que la probabilidad de que una muestra aleatoria de 100 secuencias de nuestro espacio de secuencias tenga una media del porcentaje de identidad mayor al 30% es igual al 45%.

2) probabilidad de que la media de una muestra aleatoria de 100 secuencias sea mayor al 50%

$$P\{\bar{X} > 0.50\} = P\left\{Z > \frac{0.50 - 0.29}{\frac{0.26}{\sqrt{100}}}\right\}$$

$$= P\{Z > 7.81\}$$

$$= 1 - P\{Z \leq 7.81\}$$

$$= 1 - 0.99999 \ldots$$

$$\approx 0$$

Por otra parte, se obtuvo que la probabilidad de que el promedio del porcentaje de identidad de una muestra aleatoria de 100 secuencias del espacio de secuencias sea mayor al 50% es prácticamente nula.

Discusión de Resultados

El código computacional desarrollado en esta investigación (Figura 1) es capaz de explorar de manera parcial un espacio de secuencias de tamaño 4^{150} a través de una caminata aleatoria, por lo que refleja el comportamiento que tienen las mutaciones puntuales aleatorias en el cambio de las bases nitrogenadas que conforman al ARN. Según algunos modelos de paisajes moleculares de fitness, la exploración del espacio de secuencias conducida por las mutaciones tiende hacia los óptimos del espacio que reflejan las secuencias que poseen mayor fitness, incorporando así el proceso de selección positiva en la evolución de las secuencias (Blanco et al., 2019) ; esta suposición coincide con el resultado de haber encontrado identidades con diferentes tipos de proteínas presentes en la naturaleza (Figura 3), ya que esto muestra como el proceso evolutivo conducido por la caminata aleatoria del código obtiene como resultado secuencias similares que poseen estructuras y funcionalidades complejas que se supondrían han llegado a un óptimo de alto fitness gracias a su historia evolutiva, por lo que la aleatoriedad sería una estrategia de exploración que permitiría encontrar las secuencias con mayor fitness de manera más eficiente. Este comportamiento biológico se ha traslado al desarrollo de algoritmos de optimización en donde se busca minimizar una función a través de la exploración de los diferentes valores, que pueden ser entradas de la función y evaluando el resultado en busca del valor mínimo, en este aspecto se muestra que la exploración estocástica de este espacio abstracto es más eficiente en la optimización (Gardner, 1984).

Hay que tener en cuenta las limitaciones en cuanto a la descripción biológica del código computacional, ya que este toma a la marcha aleatoria entre a,u,c y g con igual probabilidad, por lo que presupone que las mutaciones son eventos aleatorios completamente equiprobables pero en los sistemas biológicos reales se ha evidenciado que las mutaciones son sesgadas, cuya característica es el comportamiento aleatorio no uniforme; hay sustituciones de bases nitrogenadas que son más probables (Stoltzfus y Yampolsky, 2009), además de que se ha relacionado estos sesgos mutacionales con que existan zonas ricas en guaninas y citocinas en algunas partes de la secuencia (Palazzo y Kang, 2020). Secuencias obtenidas en este estudio y cualquieras otras que se obtengan a partir del código diseñado son incapaces de mostrar estas características observadas en proteínas nativas como secciones ricas en solo un tipo de base nitrogenada, ya que como consecuencia de la ley de los grandes números las bases nitrogenadas de las secuencias siempre tenderán a poseer una proporción de alrededor del 25% de aparición de cada una; esto refleja también que el código no es capaz de explorar todo el espacio de 4^{150} secuencias, por lo que hay zonas del espacio de secuencias que son inaccesibles para el código computacional.

Solo muy pocas de las secuencias obtenidas por el código diseñado son completas en el sentido de que no presentan codones de parada, siendo la gran mayoría de secuencias obtenidas incompletas especialmente con 2 o 3 codones de parada (Figura 2); esto mostraría la dificultad de generar secuencias multidominios *de novo*, ya que según lo obtenido en la muestra aleatoria es más probable que se encuentren secuencias incompletas en el espacio de secuencias. Estas observaciones están en concordancia con el planteamiento de la composición modular de las proteínas, ya que según este, las proteínas se forman por reordenamientos de módulos previamente existentes que ya han demostrado ser funcionales (Bornberg-Bauer et al., 2005),

siendo la reutilización la primera fuente en la formación de las proteínas presentes debido a que la mayoría de secuencias de proteínas que se pueden formar aleatoriamente *de novo* presentan codones de parada prematuros afectando posiblemente el plegamiento y por ende la función de estas proteínas, aunque las secuencias en su totalidad son incompletas se encontró porcentajes de identidad significativos en zonas de estas secuencias que corresponderían a dominios proteínicos lo que demuestra que sería más conveniente obtener estos dominios de *novo* en vez de las secuencias completas y después reorganizarlos para formar proteínas más complejas.

Un gran porcentaje de las secuencias generadas no obtuvieron porcentaje de identidad (Figura 3), esto se pudo haber dado porque estas secuencias son dominios huérfanos que no presentan homología conocida, este tipo de dominios están caracterizados por una distribución no aleatoria de bases nitrogenadas (Ekman et al., 2005), pero en nuestro caso las secuencias obtenidas por el código presentan una distribución aleatoria de bases nitrogenadas, por lo que habría que evaluar otros criterios como si hay presencia de estructuras secundarias en las secuencias que no obtuvieron porcentaje de identidad para poder definir si realmente se tratan de dominios huérfanos (Brown et al., 2002) o simplemente son secuencias que no se encuentran en los sistemas biológicos debido a que nunca surgieron en la historia evolutiva o que se han perdido a causa de los eventos de extinción masiva. Hay que tener en cuenta que, aunque estas secuencias no tengan identidad con secuencias que ocurren en la naturaleza no se puede rechazar la posibilidad de que estas secuencias sean capaces de desarrollar estructuras tridimensionales y por ende capaces de desempeñar funciones; esto se debe a que se ha demostrado mediante experimentos *in vitro* que secuencias de aminoácidos formadas aleatoriamente pueden generar estructuras y realizar funciones inclusive sin presentar similitud con secuencias caracterizadas en los organismos vivos (Keefe y Szostak, 2001), además, las secuencias obtenidas por el código

pseudoaleatorio se podrían considerar como secuencias cortas de marcos de lectura abiertos (sORFs), ya que en el código no se programó explícitamente que la secuencia de bases nitrogenadas empezara en un codón de inicio y terminara en un codón de parada, por lo que en muchos casos se obtuvieron pequeñas cadenas polipeptídicas dentro de la misma secuencia separadas por codones de parada. Interesantemente se ha evidenciado que los sORFs producen pequeños péptidos de menos de 100 aminoácidos con diversas funciones biológicas presentes en todos los organismos vivos, pero que debido a que desde hace muy poco tiempo se empezó a tomar en cuenta la importancia de estas proteínas pequeñas muchas de estas carecen de anotaciones en las bases de datos bioinformáticas como por ejemplo las usadas en este estudio (Olexiouk et al., 2018), esto podría ser otra de las razones por las cuales casi la mitad de las secuencias obtenidas no presentaron resultados de homología en el BLAST.

La mayoría de las secuencias generadas por este código presentaron porcentaje de identidad con algún tipo de proteína (Figura 3), todas estas secuencias superaron el 30% de identidad (Figura 4), que es un criterio para definir si dos secuencias tienen el mismo origen evolutivo (Gómez et al., 2011); por lo tanto, el código computacional pseudoaleatorio fue capaz de formar secuencias homólogas a proteínas caracterizadas en la naturaleza. Como se encontró homología con diversos dominios proteicos, puede que estos dominios que en la actualidad conforman al repertorio finito de los modulo que disponen las proteínas, fueran formados en un origen de manera aleatoria, ya que el código fue capaz de formar dominios homólogos *de novo* de manera estocástica.

Se considera que para evaluar el origen estocástico de las proteínas hay que demostrar si proteínas formadas aleatoriamente *de novo* son capaces de plegarse y de desempeñar funciones definidas (Tong et al., 2021), el análisis de probabilidad realizado evidencia que en una muestra

aleatoria de 100 secuencias formadas estocásticamente hay una probabilidad del 45% de que la media del porcentaje de identidad sea mayor al 30%, esto quiere decir, qué en algunas muestras de 100 secuencias se encontrarán secuencias aleatorias que presenten homología evolutiva con proteínas encontradas en la naturaleza, por lo tanto habrán secuencias formadas aleatoriamente con la capacidad de plegarse tridimensionalmente y de desarrollar funciones específicas similares a las proteínas homólogas naturales. Estos resultados coinciden con diferentes estudios in vitro en donde se usaban secuencias polipeptídicas provenientes de librerías de proteínas aleatorias. Se ha mostrado que secuencias formadas solamente por los aminoácidos glutamina, leucina y arginina de manera aleatoria y expresadas en células de *E. Coli* son capaces de formar estructuras secundarias y de mostrar comportamientos encontrados en proteínas naturales como la resistencia de denaturación bajo ciertas condiciones de estrés (Davidson y Sauer, 1994). Otros estudios mostraron que secuencias aleatorias formadas solo con algunos de los 20 aminoácidos o usando todos los 20 tienen comportamientos presentados por proteínas naturales como la solubilidad (Prijambada et al., 1996; Doi et al., 2005); en otro ensayo experimental usando una librería de secuencias aleatorias con una longitud de 50, encontraron que el 20% de las proteínas formadas por estas secuencias mostraban plegamientos (Chiarabelli, 2006), lo que concuerda con los resultados obtenidos por esta investigación ya que en este estudio se trabajó con secuencias de 50 aminoácidos y en estas secuencias se encontró de manera indirecta que un porcentaje de ellas (59 %) eran capaces de plegarse (ya que presentaron homología evolutiva con proteínas naturales), por lo que aunque 50 aminoácidos de longitud parecen muy pocos es posible formar dominios y motivos de manera aleatoria con propiedades estructurales complejas. En general estos estudios muestran que proteínas sin historia evolutiva generadas de *novo* aleatoriamente presentan propiedades estructurales similares a las encontradas en las proteínas naturales,

indicando así, que el desarrollo de estas propiedades no proviene de la evolución si no que es una característica propia derivadas de las propiedades fisicoquímicas de las cadenas de aminoácidos (LaBean et al., 2011).

Para evaluar la funcionalidad de secuencias polipeptídicas generadas aleatoriamente se han realizado estudios de selección dirigida in vitro, en un estudio que usó una librería de secuencias aleatorias de 80 aminoácidos de tamaño 6×10^{12} se estimó que $\approx 1\ en\ 10^{11}$ proteínas desarrollaban sitios activos con la capacidad de unirse al ATP, interesantemente estas proteínas no mostraban características estructurales similares a las proteínas naturales que desempeñan esta función, de esta manera concluyeron que las proteínas funcionales son lo suficientemente comunes para ser encontradas a través de la exploración estocástica del espacio de secuencias (Keefe y Szostak, 2001), coincidiendo con los resultados de esta investigación en donde se obtuvo que en un espacio de secuencias inclusive más grande compuesto solamente por secuencias aleatorias se podían encontrar en una muestra relativamente pequeña secuencias con funciones determinadas (Figura 3); por otra parte, en el estudio de Knopp et al. (2019) generaron sORFs a partir de secuencias aleatoria de nucleótidos, estas secuencias expresaban proteínas de 50 aminoácidos o menos; estos autores encontraron que algunas de estas proteínas de tamaño reducido al expresarse en células de *E. choli* eran capaces de conferirles resistencia a antibióticos, ya que tenían la capacidad de introducirse en la membrana celular y de esta manera afectaban la polarización celular, aunque este efecto puede ser deletéreo en muchos contextos, específicamente en ambientes con alto estrés inducido por la presencia de antibióticos, bajo pH, o bajas temperaturas; el cambio en la polarización membranal puede ser beneficioso y por ende seleccionado evolutivamente. Este estudio demuestra que péptidos pequeños producidos por secuencias aleatorias pueden ser transmembranales, esto coincide con los resultados obtenidos en

esta investigación ya que se encontró homología evolutiva con transportadores transmembranales que permiten el flujo de iones como el sodio que son fundamentales para generar el potencial de acción y por ende tienen la capacidad de afectar la polarización membranal.

Usando técnicas de biología sintética y diseño inteligente se ha logrado recientemente construir proteínas *de novo* con funciones fundamentales para la vida; por ejemplo, se han logrado formar enzimas como ATPasas, en muchos casos estas enzimas generadas de novo tiene la característica de que no comparten secuencias y estructuras con las enzimas naturales que cumplen estas mismas funciones (Donnelly et al. 2018;Wang y Hecht, 2020); a diferencia de estos reportes, en esta investigación se encontró que de manera aleatoria también se pueden formar secuencias *de novo* que desempeñan funciones enzimáticas fundamentales para la supervivencia de los sistemas biológicos, este es el caso de diferentes factores que componen a la ARN y ADN polimerasa, quinasas, hidrogenasas, etc. Hay que tener en cuenta que este estudio tiene la limitación de asignar estructuras y funciones a las secuencias generadas aleatoriamente solo con base en la identificación de homologías, por lo tanto es imposible para esta investigación definir si dentro del 41% de secuencias que no tuvieron resultados en el BLAST existen polipéptidos capaces de desempeñar funciones biológicas pero cuya estructura e identidad de secuencia es muy diferente a las proteínas naturales que desempeñan dichas funciones como sucede con las proteínas *de novo* diseñas de manera determinista a través de las tecnologías desarrolladas por la biología sintética.

Aunque se encontraron porcentajes de identidad relativamente altos (Figura 4), en un estudio que uso herramientas bioinformáticas para comparar secuencias aleatorias con proteínas naturales encontraron solo concordancias de baja significancia en los alineamientos al realizar el BLAST, por lo que no tomaron en cuenta los resultados de similaridad obtenidos (Tretyachenko

et al., 2017); esto puede estar relacionado con el tamaño de las secuencias analizadas, ya que nuestra investigación se realizó con secuencias cortas en comparación con las secuencias de 109 residuos con las que se hizo ese estudio, a pesar de esto en ese estudio encontraron que el contenido de estructuras secundarias de las secuencias aleatorias era muy similar al encontrado en las proteínas naturales.

Cuando se calculó la probabilidad de que el promedio de identidad de una muestra de 100 secuencias sea mayor al 50% se obtuvo una probabilidad nula, esto puede ser debido a que en cada muestra aleatoria deben haber secuencias que no presenten porcentaje de identidad como lo corrobora la muestra obtenida(Figura 3), esto generará que el promedio de identidad se vea drásticamente afectado disminuyendo su valor, por lo que encontrar muestras aleatorias de 100 secuencias generadas estocásticamente que tengan un promedio de identidad mayor al 50% es prácticamente imposible. Este hecho refleja de alguna manera que el espacio de secuencias definido es este trabajo esta permeado por secuencias aleatorias muy diferentes a las secuencias caracterizadas que se encuentran en la naturaleza y en donde algunas de estas posiblemente no presenten funcionalidad biológica.

Conclusiones

El código computacional es capaz de generar una caminata aleatoria en un espacio de secuencias para obtener secuencias de ARN de manera pseudoaleatoria de esta manera simulando en proceso de evolución molecular conducido por mutaciones puntuales. Al traducir las secuencias obtenidas a través del código de ARN a aminoácidos se encontraron secuencias que en su mayoría contenían codones de parada, pero que aun así presentaron homología con diferentes tipos de proteínas (enzimas, transportadores, proteínas de unión, etc.). A través del análisis probabilístico de los porcentajes de identidad obtenidos se obtuvo que en una muestra

aleatoria de 100 secuencias formadas estocásticamente hay una probabilidad del 45% de que la media del porcentaje de similitud sea mayor al 30%, demostrando así que es probable encontrar secuencias que presenten homología en algunas muestras aleatoria; por lo que es posible que en un origen las secuencias proteicas se hayan generado de manera aleatoria ya existen secuencias aleatorias capaces de plegarse y desarrollar funcionalidades específicas.

Referencias

Blanco, C., Janzen, E., Pressman, A., Saha, R., & Chen, I. A. (2019). Molecular Fitness Landscapes from High-Coverage Sequence Profiling. *Annual Review of Biophysics*, *48*, 1–18. https://doi.org/10.1146/annurev-biophys-052118-115333

Bornberg-Bauer, E., Beaussart, F., Kummerfeld, S. K., Teichmann, S. A., & Weiner, J. (2005). The evolution of domain arrangements in proteins and interaction networks. *Cellular and Molecular Life Sciences*, *62*(4), 435–445. https://doi.org/10.1007/s00018-004-4416-1

Bornberg-Bauer, E., Huylmans, A. K., & Sikosek, T. (2010). How do new proteins arise? *Current Opinion in Structural Biology*, *20*(3), 390–396. https://doi.org/10.1016/j.sbi.2010.02.005

Brown, C. J., Takayama, S., Campen, A. M., Vise, P., Marshall, T. W., Oldfield, C. J., Williams, C. J., & Keith Dunker, A. (2002). Evolutionary rate heterogeneity in proteins with long disordered regions. *Journal of Molecular Evolution*, *55*(1), 104–110. https://doi.org/10.1007/s00239-001-2309-6

Chess, A. (2013). Random and non-random monoallelic expression. *Neuropsychopharmacology*, *38*(1), 55–61. https://doi.org/10.1038/npp.2012.85

Chiarabelli, C., Vrijbloed, J., Thomas, R., & Luisi, P. (2006). Investigation of de novo totally random biosequences, Part I: A general method for in vitro selection of folded domains from a random polypeptide library displayed on phage. *Chem Biodivers*, *3*, 827-839. https://doi.org/ 10.1002/cbdv.200690087

Davidson, A. R., & Sauer, R. T. (1994). Folded proteins occur frequently in libraries of random amino acid sequences. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(6), 2146–2150. https://doi.org/10.1073/pnas.91.6.2146

Delport, W., Scheffler, K., Botha, G., Gravenor, M. B., Muse, S. V., & Kosakovsky Pond, S. L. K. (2010). CodonTest: Modeling amino acid substitution preferences in coding sequences. *PLoS Computational Biology*, *6*(8). https://doi.org/10.1371/journal.pcbi.1000885

Dessalles, R. (2017). *Stochastic models for protein production: the impact of autoregulation, cell cycle and protein production interactions on gene expression* [doctoral thesis]. Universidad de Paris-Saclay

Doi, N., Kakukawa, K., Oishi, Y., & Yanagawa, H. (2005). High solubility of random-sequence proteins consisting of five kinds of primitive amino acids. *Protein Engineering, Design and Selection*, *18*(6), 279–284. https://doi.org/10.1093/protein/gzi034

Donnelly, A. E., Murphy, G. S., Digianantonio, K. M., & Hecht, M. H. (2018). A de novo enzyme catalyzes a life-sustaining reaction in Escherichia coli. *Nature Chemical Biology*, *14*(3), 253–255. https://doi.org/10.1038/nchembio.2550

Ekman, D., Björklund, Å. K., Frey-Skött, J., & Elofsson, A. (2005). Multi-domain proteins in the three kingdoms of life: Orphan domains and other unassigned regions. *Journal of Molecular Biology*, *348*(1), 231–243. https://doi.org/10.1016/j.jmb.2005.02.007

Fragata, I., Blanckaert, A., Dias Louro, M. A., Liberles, D. A., & Bank, C. (2019). Evolution in the light of fitness landscape theory. *Trends in Ecology and Evolution*, *34*(1), 69-82. https://doi.org/10.1016/j.tree.2018.10.009

Gardner, W. A. (1984). Learning characteristics of stochastic-gradient-descent algorithms: A general study, analysis, and critique. *Signal Processing, 6*(2), 113-133. https://doi.org/10.1016/0165-1684(84)90013-6

Gómez, J., González, A., Castaño, J., & Patarroyo, M. (2011). Biología molecular: principios y aplicaciones. Fondo Editorial.

Jensen, M. H., Morris, E. J., Tran, H., Nash, M. A., & Tan, C. (2020). Stochastic ordering of complexoform protein assembly by genetic circuits. *PLoS Computational Biology, 16*(6), 1–18. https://doi.org/10.1371/journal.pcbi.1007997

Keefe, A. D., & Szostak, J. W. (2001). Functional proteins from a random-sequence library. *Nature*, *410*(6829), 715–718. https://doi.org/10.1038/35070613

Kelley, J. L., & Swanson, W. J. (2008). Positive Selection in the Human Genome: From Genome Scans to Biological Significance. *Annual Review of Genomics and Human Genetics, 9,* 143-160. https://doi.org/10.1146/annurev.genom.9.081307.164411

Kimura, M. (1991). The neutral theory of molecular evolution: A review of recent evidence. In *The Japanese Journal of Genetics*, *66*(4), 367–386. https://doi.org/10.1266/jjg.66.367

Knopp, M., Gudmundsdottir, J., Nilsson, T., König, F., Warsi, O., Rajer, F., Ädelroth, P., & Andersson, D. (2019). *De novo* emergence of peptides that confer antibiotic resistance. *mBio,10*(3), 1-15. https://doi.org/10.1128/mBio .00837-19.

LaBean, T. H., Butt, T. R., Kauffman, S. A., & Schultes, E. A. (2011). Protein folding absent selection. *Genes*, *2*(3), 608–626. https://doi.org/10.3390/genes2030608

Lodish, H., Berk, A., Kaiser, C., Krieger, M., Scott, M., Bretscher, A., Ploegh, H., & Matsudaira, P. (2008). *Molecular cell biology.* W. H. Freeman and Company.

Moore, A. D., Björklund, Å. K., Ekman, D., Bornberg-Bauer, E., & Elofsson, A. (2008). Arrangements in the modular evolution of proteins. *Trends in Biochemical Sciences, 33*(9), 444–451. https://doi.org/10.1016/j.tibs.2008.05.008

Nei, M., Suzuki, Y., & Nozawa, M. (2010). The Neutral Theory of Molecular Evolution in the Genomic Era. *Annual Review of Genomics and Human Genetics, 11*, 265-289. https://doi.org/10.1146/annurev-genom-082908-150129

Olexiouk, V., Van Criekinge, W., & Menschaert, G. (2018). An update on sORFs.org: A repository of small ORFs identified by ribosome profiling. *Nucleic Acids Research, 46*(D1), D497–D502. https://doi.org/10.1093/nar/gkx1130

Palazzo, A., Kang, Y. (2020). GC-content biases in protein-coding genes act as an "mRNA identity" feature for nuclear export. *BioEssays, 43*(2), 1-11. https://doi.org/10.1002/bies.202000197

Prijambada, I., Yomo, T., Tanaka, F., Kawama, T., Yamamoto, K., Hasegawa, A., Shima, Y., Negoro, S., & Urabe, I. (1996) Solubility of artificial proteins with random sequences. *FEBS Lett, 382*, 21-25. https://doi.org/10.1016/0014-5793(96)00123-8

Pasek, S., Risler, J. L., & Brézellec, P. (2006). Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics, 22*(12), 1418–1423. https://doi.org/10.1093/bioinformatics/btl135

Rice, P., Longden, I., Bleasby, A. (2000). EMBOSS: The European molecular biology open software suite. *Trends in Genetics*, *16***,** 276-277. https://doi.org/10.1016/S0168-9525(00)02024-2

Saiz, L., & Vilar, J. M. G. (2006). Stochastic dynamics of macromolecular-assembly networks. *Molecular Systems Biology*, *2*, 1–11. https://doi.org/10.1038/msb4100061

Stoltzfus, A., & Yampolsky, L. Y. (2009). Climbing mount probable: Mutation as a cause of nonrandomness in evolution. *Journal of Heredity*, *100*(5), 637–647. https://doi.org/10.1093/jhered/esp048

The Uniprot Consortium. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*, 480–489. https://doi.org/10.1093/nar/gkaa1100

Tong, C. L., Lee, K., & Seelig, B. (2021). De novo proteins from random sequences through in vitro evolution. *Current Opinion in Structural Biology*, *68*, 129–134. https://doi.org/10.1016/j.sbi.2020.12.014

Tretyachenko, V., Vymětal, J., Bednárová, L., Kopecký, V., Hofbauerová, K., Jindrová, H., Hubálek, M., Souček, R., Konvalinka, J., Vondrášek, J., & Hlouchová, K. (2017). Random protein sequences can form defined secondary structures and are well-tolerated in vivo. *Scientific Reports*, *7*(1), 2–10. https://doi.org/10.1038/s41598-017-15635-8

Trifonov, E. N., & Frenkel, Z. M. (2009). Evolution of protein modularity. *Current Opinion in Structural Biology*, *19*(3), 335–340. https://doi.org/10.1016/j.sbi.2009.03.007

Van Rossum, G., & Drake, F. L. Eds. (2011). *Python language reference manual.* Network theory Ltd.

Wang, M. S., & Hecht, M. H. (2020). A Completely de Novo ATPase from Combinatorial Protein Design. *Journal of the American Chemical Society*, *142*(36), 15230–15234. https://doi.org/10.1021/jacs.0c02954

Weiner, J., Beaussart, F., & Bornberg-Bauer, E. (2006). Domain deletions and substitutions in the modular protein evolution. *FEBS Journal*, *273*(9), 2037–2047. https://doi.org/10.1111/j.1742-4658.2006.05220

Apéndice A

Tabla 1: 100 Secuencias de ARN Mensajero Obtenidas a Través del Código Computacional Pseudoaleatorio

ARNm
auugaggcgcgucucacgcuaggacuugacucauagcugcucuaugccgguccugacuugucugcgccgacuacucgucucgcuaaguugggguagggguaggaguuagcgcuuuauuccgaccuacaacugcgcguacgcgcaucuaguaa
gaauacauaagagagucggggccgguaucuagagcccauuuagucaauauuuuggacgacgcuauuacuuggcucccccacgaaaccuucaaaggccuuggcgggguaaacgcgaacuauaacauguggguaccggugacgaaaucaagacca
acaacugaaucgggcuaaucaagcggacuucucgaauacgagcaacgucacuaucauuaauaugucgaugccaguuugcgucccgauugagcuuggccuguauuauuaucagagcuggcauaugagggcuccucccguguaccauucuccc
acuauaaggccaucauacuauguauggaggggagcgauuggaggacgguuacccacacuugcacccguuaagagcaaagcccaugcguuuuggaauguacaagguacgccugauucugccauccgccauagucccaauaaucuggccuuga
gacaacaugacgcacgcguauuuagagucuugaucgaucguccauucggucgcuuuugcauaagcugugaauuagugauucagaucauaauuguggguggaguccaccccagaugucacucccccuaugugucguuggccccuuccgaaacg
gucgauccuuucgauuacaugauuauagaaagacucauacaagggauuuacggacgggugagcgggaagaucaaggcccguaacucauguacagcuucuguugucgaccaggcgaauuacgggugugggcaauggugagguuguauggga
uacaagagcaccauuuccgguucacuacaucauucccagaaaaacccaacauagcucaagucucaccucugcacuacgcccaucgccgcuuuuguaauugcgguggaucgacucaaaaaucucaaaccaaggccuuaaugcggacaucag
gagauccaaaaucuacggcgcuagaauggagaugccuuagagcaacgccgcauaaaggauucacuucgcaacccuuuagcccaguguaucguuaagucuugcguauguuguuguuggccgucuguuaagaugaacccgguccagauuc
aagucauuagucggauucguuuggacgcgauuauuuuguaggcuggacaacgucccgguggcacucagugcuacgcuagcucguccggcccagcugcgaagaauggcacuaccugcguagaagcuacagacaucagugaaagccuccuga
uggggauccauacucuagccgaagcgguauugaugggcgggauaguggagucucagucuauaaauccgacguacgagguugccgcgugguguacgggacugcgacaauaugcgcacggagccucgcaucuccacuugaaucaacagagaga
cagccuguccggucugcccgaggguaccacaucaaaaaacaacggaagcaaucacacaaccuuuugaugguugccaguuaaugcuucgacaacgcuucuuguaccuaaugauaccccccagucgacaagauguaaaggguggcaccgcau
uucagaccucccacaagggugcugguuagcagcguuaauauggccggaucggaugcauccaccaauugaaucgcuagcgagaacccgauugcuuggugcugucuaacucacccguaaucggccuaugcccaauuggggcaccccuaggaga
acaugccuccaaauuauuuauuccaugauuuagcuucgccugaacuucuacagacgcuccccccacugcuuuacuagauggacaucugccuggagagguucgaguaaucugacaacccauaccgggggagaaccauuagccccccuguuccu
agaauuuaguuguagcggcauucuucgagccgcaaaguucgcccggccggaauggcuugccuguaauuuacgugaagacguaagauaaaggauucauagcaucagaugcuccugugaagugcucccccuccuaccggcauuacgcucuaa
cccacucucaacggcaggugagguaaaauuggcaaagacaguggggucguugucaacggaucuagcuggcgcgggguuggagcaacaaaaauggguaagugucgaguacaugaacuugacacacgacccuucucauucgaacggacagcacua
cugaacauucuaaccuuucgcguuuaggucgugcaaaguaacuacgagcgaugcuuuaucauacuguuucguccgcgcauggggucuucgggacaucgcgaccgcaacaccagaacccgaauaucagauauugaaugaagacgaugccuc
gagucggcggggguguacuagggaccuaguccucggacuuugcggucccucgcaaguuuuuagauguaaugaagacgcauucccaacgguguaugcguggcugcagaugagccgcuaucacaugccuuuugguauggaggaaggaccucgu
gcaauccgcagauggaacacucaaaaacauacgcgcaugccgguuccgcggguagagacuauuucuuauucccuuagagcgccuggcuuggguuuuccgagugguaaugauacgaggugaaggcgccacgcauaguacgggguuugcgggc
ccuagaggcgaaggggguucacucaaacggcgugaugcuucucuacuggcaguauaaauacaguacggcaguccagauuccggccacacguccgccagggcucuucgauccgaguuaucccauaauugcuuaugguccgcucgccaccac
gugagcugucgcgcuuucuaggguuuucaacaucgucguuaagaugcgucuagccuacaucucaagagaagauuucaugucccaccaccaauugggucgggguucaaagcuagaggcauugccaaguaugaagcgugucgagagcugagaau
ccguggaucuagguccaggcgacgaaucaagacggguacggucgacuauaguaauaccuauuucaugcaugccuaguagcaaauaauuaacacaguaagcgaccguucgaggguagugagucccgcucaggcaaauuccgauccuucgca
ugggagggugcagaugaggucaccaugcaacaaggugauaugaaaggcuggacauaaugguuacacaugugauuugggugagucggacaucgagcuuacccuauaucagcuaguuagccagcgcauccucauuggcagggagguaccacu
aacgaguggguacgucacugccagacgacacgugggcugggaacgcuuaaauuccagaauccuacaaucggcucgugucguacgacgauauugcggguugagcuaccccuuuuugcguucuagagcuccugcacgguacacaguggugcca
ggcacucauauacuuaagcgcucaaccucacugacaguuucagcacugacgccgaggaguagccguggucucggagaagacccauccgguuguacggaaguacagguagauucguguggagcuccgagauaguauguagaacgcgacggu

uucaacuuaucggcgaauggcuuugaguaagaaaagcucgaaagcgauuuguaggaggcaaaauacaaugggggccccccaucucucauucgaagcuacuaguuuuug gacauugguccuaccagaggaucuaaccgccgaguccggaggu
uauagacugguucgaacacucgucuauuggggucucgucgggggugagaugccuaaggaccagugagaaucaaucgcuggcuuauugcuuacggugcucguacuggcc cauggcuugaaauacagcagugcaggggauguuugcguugcucc
guaugacaauacauaauaggaacucaugacuaaucagagcccgacgcauuacgaucaauauauuugcauugaaacuguucgaacacggcgacaggccaggugccauuu ggacagacgauaauacaggguuuguucugagcuccauauagu
cggaguaaguagguugccugggacggcaauaugggguagcugaguucaccgaggucuccuaaucguaguggggugagcggauacccuacguccacaggcuucaggga gaaagacuuuuaaaccuagugacuuaaagaaaagaugcuggcca
aagguauaauggaaucggauaacuucguggccgcaaauuccggacgccccgcgcuugcucucgaccauuacgagcgccaggauggaguacaggugcauacaggucaa ugauaggaaaugcagucgcaugagcaaguccgagaacucuaag
accugugggacauacgucugcaucccccgcauacugugacucacagacaccuguuucagagguuuuuagcauuacuuuauuucucgcggaugugcccuuuaccaccg cucggugaauuggugcaagugugccaguuucagaaccaagcga
gguagggcgucguuagucugacaccuguaucgcucuagcacuucgugauggaccaacgucauaggauuauuuaaucagcccaucccucgugaccuaccaucccucg aaagcuaauaauauaucuuugcugccgaagaggugaacaucug
ugcgcuguccauagacccgcaaucuagcccuuaagaggcggccgccgcaccgauacucuacgaaucucagggaagguuacaugcugaacagacaguucgcagucuaua ccauauguggcacgugugauuuggguguuagcguaacaacagg
cuccauauucuccacccaguauugccuuuaacgagcucaacuuauggacgugcuucagcagcuucggucuggggggccacgguuccaagcccggccacgucucaagg acaacuaagaguuucgucgaagcuaccuugaucauggagauag
gccacggauacacuagcggacccucugccccgucuccauaguucuugcguugcgagucagcgccuuuuacgccauuauaucucuuuaaagauuaguacgccugcgca cguaucgggaauucgccacuagccgccaguucaccgcgaucaa
uaccuggugugacucuccucucagagagaacgcagaaauaggaccucauccaaagguauuggcacgguacuaaguguucgagggagucgcuuucaggcgcuuaucgc gggguuacccuguaugaggaucuacaugaagucaugcggcugc
augggguaacgucgcacacucgcuagacgaaugcucugaucugccaccagagucgcgugcagaugggacggcuuugguagaacccgaggcggacauccuagcuguucg ugacaagaguugcauugauaccgccagcaugaugaggauacgg
ucaacucagaugacgauccguuccagcuugcagcauagcgggggaaugacaaauaacacacaccgucccggggaacccgaacguucagcaaaagacauuuggauugcc ccgagaguacccacucucccggucauuauauucgcgaugagc
guauaaacaaguccuacacugaagcuggcacucggagagcauaagccgaaagaagggccagccggcuauucuuccuacucugaguuugagaccuuacaugguccucu uuacucucucuagcgcacuguucaagucuugggaauucuucgg
uuggucgauagguucgagaaccacaugcugggccaugucgcaugaugauauaauaggaguuggaagcgguuaaaugggcguagguuggcgguucaauaaggggaaa ugaaaucgagaugauagggagacaggguaucguuaaauuauuaa
cuuuaaucguccacaaacccaccuugacacagcgguaucgaaacccuggauguaaggauaaucuacuugauccgucuagcuccgccagugcuaguugaccgcagcggg ggccuuaagaguggucguuagacuccgugcgccaugcuggga
cugccgccuaaauggaagggcucucaaagaggaaccaguugcuuaguccucggugaccaccucaugucauagcauugguucuguaacuuugcaggcccaucgcauua gacugcuauagacugaauguuaccguacaguaauaaaguggcu
uugaugcggaccuacugcggaggcugugucguggauaacgagggcuauccggcuauucugcccgcgcuacauaucugaggcccgagggacuuuuaagucugaaggc auaguugcggaccgugccauaguuauggucaccguacgaacaau
uacaucugcgucaaaagaauucucagguaacgugguaacaccagcguccaaugacaagccuacuaaggcaacccauccuguaagcuguauuaccgaauccuacuaaag cgguccgagcccguuucagucgguuugccggaggguccauac
auguaaauugacagccggcggauuuuauuaacauacagcugcacgagacagcgugaaaaaaaacagcuggcucggccccuugugucauucggcauggcgucugaucc ccaaucagcgcggcguacaguggcaaaagguauucccaauag
agcgcgggucuaauucgccagagagguuccucucgagagugauacaaucaaguagcauccggucuguuaaaggggacuaaggacaacuuaauaaaugggugcaacggag ccccccuacaagcugaagcggcuaucgagcuaccaacuauuaau
ggcuaauagcaagagcucacaaaacacgacuggcuuuauggcuuuggcgaacgcuuaaaaacucccacgauggucaguuuacuucaucuuggggaugggccggcgucg caagaguucguacggucucuuaugaguaugguuggcccauccc
auacuugugauggcgcgcuaucuuucuacuucguucgcgcacgacgccuauaaauauacccauagaccucgcaauucuuguacaccaggagucaucggaagccagccgc cauaaagucaugguguuuacgacaaaccuuuuuucgcgucgc
gcauagacuggagggauuguacaaacgcuguaagaacugaaccgugccugaggggucugugcgauucuagugucucgccccucguaucgaguguuacggugguugc auuacaaaugguccaguggugaugcgcucucaagggcccgagaa
cgcacacauaaccggaucauauugcucacaacgcauaccuuacguaucgagcgcauagcgcaguuucaguguuucguaaucuuuagcaacauucggacauuaugaccc auucgagucgacaacacuauggagacgagcgcaacucaggcu
acaauaccagaaagauaacagucgagucaagcgugucggcccgccgagcguaugaccaucaggauauacuuuuggccuagacuucguacgauccagaaccucaaaucg ccgcucauagacaggacggggucuguagcaacacaugaggag
cauuucuguucacugcacgcuuacggaaugcucccagcgaaucguaugcauaggcgccgauggagaaaaaacuauaaggaauaacugaagaaggacuaauccgcgucu acagauuuacccuugccauccuagcgcuaagccuggaggaac
ugggcauacgucuuugggaaucagaaacgaugccuuagaucccuucacgggugauauggcuauccugccggacgagcgguauaauugugucucuaaaugcaauacu uaauguccgccuucuagcgcgcaaagucucuucacacagucaag
gacuagacagagauaaauaggcugacgaguguagauguuucuagaauaaguagagucccgugauggcuaucccaugacguaacauucacgcguccgcaauaucuagcg gaaucaucccaaaucaccagcucucacgguuguuuucccauug
aguuccaccgccgugaaaagaguuuccgggucggcacuaugguacucauagcacgcaguaggcgcuaugauuacccuacuauggagacuauaugccgguaccguccg ggauguuaaguuuuguccggccugaguaauguauguuggcguc

ucauccgugccacgucucaaaacaggucacucaacuggagcacuacauuaccucgauauucgcgaauauaucccggucuacguauccugaugaacuaugcaacgucga cggcaagauuagauagagccguuaucgaagaauuuuaucccuc
gacccagccuaaaccuaucgcgcucguaggauucguaagcugaaaagcgcuuuuguuuagcagcccggugaagcucaaauccggaccuaguuaaggguuuaaugauuc uaucgggggguguacccguccaacgacucguggcagggaagguu
ucaacuacugaacgugcugcucuggguaacuucgucacuucguaccauaaaauacaaacgcacuccuacccguacugguaaugggcgaagauugacaccuaaagucgac ucucgccauaccgggcacacguuuuucccgagucaggugugu
ugaaaucauagaaacgccgcuaugcugucuuacgacuaaguaaaaacgaccgccguggauuaccuuugauuuaaucaauaugauaugagacuaagauaggcugagcg cacccaucugaauggucuauccgaaauauaguguuggcacgcc
augcauccgccaguaguccucgacuuuccauucuugucgagcgcccaucgguaccuaaaccgagaagcgauccugccuuaaacccgcggagccagcgaguacuggugc acuccuaggauagcgucagaggggauugugaugaaucucagu
gggcacguccgcagccggcaacuagcuccauguuaccacguguuggcaaaaaagcauguacuuucauggucaucgugauagcugugcugacuaucccaggaccucac acaagcccggugagucggacgucggacgaaacgaccgucgaau
gaauuuauagcaggcaaaucucagcgccaucuuaacaaaaucuauggaaccuagcaaugcuccuuuagcgcgggagggauacuggugagagguacgauagggucgau aacgugccaccuaguaauugacgaguccagaugcgccugcacg
uaaguugaagggacguuucuaccucuagguuuuucugacugugcgagaguuauucugaaaaucucaagcuuaaaccggugggaauauuucgcgcgguggaauagguca ugcugcgcggaauaugagaccgugcacagggaugccgcugccga
uaagcuaaaauaaucuaaagacccuugcugugaggauguaucccguucauaccgauauaucugcccucuuauauuucgacucccgguaaguacauucgcaauucga uggugagggaggauccuguguuucaaggagaaccggaaguuga
caacccgcgacagggcggugcuuagcuugcccgcaugucggcuaugguaaaguacguauccuuggucgaccaaauaccugcgagaguccucgcuugcugagcgccgg gaagauaaguucuucuuacuggcugauuaggagcccauucauc
aucauccgcucgucggucguaaguuggaacgaucggaacaacgaagaaaguggguuacgccccugccaaacagggagcgucuuuaauccgcccgguagguuucggaau acacccucguaggacagugcuacacuccgauguaccuguagca
uagccgagcggaaacccucuugaguccgccuuggcuuguuugucaucugcaccucaugugccaguucgaaaguuugcugagcacagcgggcgucugcgagaacugu gcugugauacaagcaugcgacacuucugcggcuguggccuggggg
uugucuauuuccauuccugucucuugcgccagcaccuggaaccucugucggcuaguagaugucccacuuuauaacgccaaauauacagguuccgucguuaguugug uuggcauagugcauuauuuuauuccagaccuaaugcauuggggc
ugaaugcgaacuagcuaacuuuauagaccugcugggaagaacuagcggcaauccgcggcaucggcuguuaaacaccucgucguggccgcgcaccccgguacuuguaua auaguugggcuacacgggaucacucaccucuaaggggaaccuc
ucaggccuuugaauuuauuuauggacccgccggaauauccaccgccacgaagggguguuggcaaaccuauaacggguagcuucagcaaucuuuacaacccgaguuua cauguacaacgacgcucauugcgcuuaacgacuguccaguaau
caggaccaguccuuuccgguccgggcacugcccgaacuagcaaucgguaaacguugcggcugcguacguauugagcuuccaguccgacacauguaguaagccgccca uauaaaugaauuaguucacgcucccuacuagacaaauauuuuc
gaaggaucucguuggacgcgguacgcuggauaaccgcaugcucugugcgccuuaaauguguaguacuggauaaaguacacgcacaagagcauaacugauuuuauuuc cagcgguaacaaaaauugcgaguccugcgaguugcaacuagacu
ugaagacuggauaucguguccgcuuuuaucugcuacuagcaucaguaacaaaagcauacgaccuggcagcuguucccagaagcacagcucguaauuaagugaaaagac auuuuguuucuacgucccuuguuagcgcugauccuaagacua
ccucgaacuuaagauaacuacagcuagucaccuagaggugacuuuaauguaaaguacagacccugauguuuguuccuuuauggaguuuggccacguagagauucacg gcaaucaccgaauuuccugcauccgcucuaacgcgaagcaaua
acccccagagucggucauccaagccgccagacuauccaagggaugaaacgccaccauccuauguauucagucguugucucccaauucccaucgcacccaugauuccucc ugccaugcaagaggugcuacgucgccaauagagaauauauu
uucgccccuagcuucccuaggcagacaauugacgccgcaaauugccaucccgcgcuagggcgugcgugcgcaggggugucgcagauacuagagcgcgcauggugagg auccuagugccccggaugucagaacguccaagaaaggguuucca
aaagcagugagagacgucgguaaccgaugaagguacauuggguuacuuuuuugcgauuauuguuacacucacuuuauagacgugaucauguauccuggucgggugcaa gccgacagccaguhugcaccuuucgucucgacgagaaauauuuc
uuucugacguguucggucuuucagaaguguaucuagcaccgcugcgcacagacccucuucuaccgguacccauagccccugaccuccucaagacuauucagcucgaa aagcggacacgaacuugagugggcgucuuccgcucgguuaaac
cccugggguaggguucccgaccggucacugaguugaugucggugcuuauguuguugccguaguggucgagauacucaaggcugccacucuguaucacauuaaauguau uguuuaguaauaugugaccgcucggcccggaccaaauaaauggg
cuuuagagacugggagguaggaugcggaucuaacagucccgugaacggaugaccccuauucuuaggaaacucuuuacaugaguccaacgacguagugaccccacuuag guagguaucuacggcucuacgguagagcuaugguagcauuu
uuaauguggugcagaucuuaguaucacauugcguaccugcaguaugcgugggcacuaacaccauggagggcaucaaugccguacgguguucucauugggccgcuuc ugguucguuggucuacguaguaccccgucugucgggguggc
aaaaggacucagaugcuggccgauggacacaagauaacggguaguacccucgccuggucggacuucgcccuacuucgcgugagugcuguucugcucgcgcaguuuuu ucuuuucgcauccaugcuuaguguauccggcgaccgcccuu
uauguaggacauagcaucccccguuuggggggaauagaucgaaggaauacauaaggcgccgccguagagaucgguucucugcaguaauauuggggguuauucauac uaaggaauagcuuaugccaaguguuuuccuuaaggauaggcg
ggaccucacgaggauccgcaguacucccuauccucgguccguccgcaacacauaaaacuauauucaaaacuccacuguuaaaagauaaguaucuugcagagucaagua agacuaguuccguaguuccuuggcgagaaguaaucccauu
gacgugacgacaaugcugagccucacagaauucuucggcaucaauggcaagaaagcucuguugcugcuagcagugaccuacgguaccaagccuagacuccagugcauc guuauguugauucuuaaccaacuaguuaagaaagauccac

cuaaugucacagccucccaguuacuuuucgccccagagcccggcguucugggcgcuaccccaggaccuugcucaauaaggcgcuuuacccgccacgcggugcacacg gcccguccagauaucuuagcuuaacguucaguuauacca
gcgaacugcgggcgacagcagcgcgugagaagaaagggcacucgcggugauaggaaagagggguguagcuuuaaccaguggcgacaccaaccccaccauuaucuca ucuugugcguuaugacuauaguuacuagcugggcuauccc
aaacacaugugccggaaccagaccccuaacggaucugucuggaaagaaaccaugaggcgacccggcaauuagcguaaaaacucuugacggcccuacacgcaaucauucc ggucugugaggccucuuguccguuaacgaggcauugcg
acagcuauauugacuaggcucauucccccaccuggggaguccuaaagaaccgaaagaggguuauacuugaaugaccagucgcaugucgucuuagcgagcuuuccgaau cuaucggcgacaaucauucgcacccuaucgagaggauuac
ccgccauguaaaaacuuuacgcauacgcacacccccugauuuuaggcauagauagagacuccuacaaucccgcgcgaaacaccucuccgguaauaaggaaccggaaugu cuacagguuugguaguugcgccuugcauuagaugccaa
ucucugggccgcgcggauagcuugccccgcagcggugauuucgcagcgggauccggagcugacauuguacaugcuuuuguuacugaauuuggcuuagcaccucgca cagauuuagucggcuacuucugacauuaaucguuaaugugu
ccuggguguagcuaacggaauacuuacgcaugaaucuucgggcuuggggcgggcguaaaggggcaacgcaauguugagcgcaccauugcagcgaauguucguagcag ugcucgaggguuucacgcuuaguacccugacagggcauuucu
uauaagcaauauaccgguucgagcuucuaucuccacccccggugcauccuucuggugugcacucgccuaaucuccgacagauaaggaauguuauuuagccugggaua ccaggcgggggucacugccauaggacccccuuauaauaca
cuuugggccaccgccacugaagcacuuuagguacaguggucgcccgaaggacgggcgcucuuuuuacucgacuggggacauaccugguggcaguagaggauggcgc aguacaauaaacugaacgguguaauaaugaacccgguagaa
ggagggaacucguuugggcuuccagaacuuaggaaaguaacucaugcucaggcuucaggucauaaauucgucgcagcgguacuugcaaugggaaugguguucgugcc guaguacccacuguauuagacauaugugccauauagacacg
ugagcacuacaugaccuuccuaggacgccgggugcgugucuugcacaaccauuaugcaguagcugaacauuauaaccccaucagacguggcccauuaugcuaagcuaau ugcugcgggucucccgacucgggugucgguaacucauagaac
ggcuucaucgccauauaggggcguuuuaucggaguucggagaguauuaagaacacgaaggguaguuauuccagcacagcacacgaggcucaugaauucgcuuccgcg auuccguuacaggcggcuauagcucuaagaaguuugucgg
agaucgggccuaacucugugagguaggcaaaacaaauucguagcgauacuguaggaauguggcuucgggcgccggauuaaacagcacucuaugaauggcacugaugg gaccgauaaagggagauguaguacauaggcucugaaugau
aaaagacucgaacaguaccuagugugagauugaacagggagaagagauugauucuguugcccuuacguaaagguacuacuagacuuuagaguagugacgccuucgaa gaaugcaggauguagacggcccgcauuaaaacaaguaugg
cagucuuaugggucuggacuaauguauucaguaauuauguaucagaauggcuccucccaauauuguuaaugggcacagcaaaacccccuuauaaaucuggugauccau uuggccccggucaaauuucgcaaggcgauuuacucaaggc
cacgugaaagcgggggcuccauuguucaugggcgaucguagagucggaggacaguuauaaagcuucgaagugaugcggaaucuagaucuuacaaccgagcguccaua ucuaaauuagaucgauugugccaugacuaucccgccagac
guauuagaagugugcggguccaucaagugauagggauggucagcgucuucagagcguuacagcuacggccgcucggacgacuuuacgcagagcguugucuuaaagu cuuauuacagcguguucggaaacuucgugacugcgcggaag

Apéndice B

Tabla 2: Resultados del Porcentaje de Identidad Obtenidos en el BLAST

secuencia primaria de aa	BLAST
IEARLTLGLDS*LLYAGPDLSAPTTRLAKLGRVGVSALF RPTTARTRI**	Putative MFS-type transporter C1271,10c (Talaromyces islandicus) 61.5%
EYIRESGPVSRAHLVNILDDAITWLPTKPSKALAGKREL* HVVPVTKSRP	Protein kinase domain-containing protein (Trametes coccinea BRFM310) 45.5%
TTESG*SSGLLEYEQRHYH*YVDASLRPIELGLYYYQSW HMRAPPVYHSP	ninguna identidad
TIRPSYYVWRGAIGGRLPTLAPVKSKAHAFWMYKVRLI LPSAIVPIIWP*	Fructose-bisphosphate aldolase (Rhizobium sp. BK251) 48.8%
DNMTHAYLES*SIVHSVAFA*AVN**FRS*LWWSPPQMS LPLVSLAPSET	AT-rich interactive domain-containing protein 3B (Thamnophis sirtalis) 75.0%
VDPFDYMIIERLIQGIYGR*AGRSRPVTHVQLLLSTRRIT GVGQW*GCMG	ninguna identidad
YKSTISGSLHHSQKNPT*LKSHLCTTPIAAFVIAVDRLKN LKPRP*CGHQ	ninguna identidad
EIQNLRR*NGDALEQRRIKDSLRNPLAQCIVKSCVCCCC GRLLR*TRSRF	Choline transporter-like protein (Strongylocentrotus purpuratus) 52.2%
KSLVGFVWTRLFCRLDNVPVALSATLARPAQLRRMALP A*KLQTSVKAS*	Glutathione S-transferase N-terminal domain-containing protein (Duganella sp. LX20W) 57.7%
WGSIL*PKRY*WAG*WSLSL*SDVRGCRVVYGTATICAR SLASPLESTER	ninguna identidad
QPVRSARGYHIKKQRKQSHNLLMVAS*CFDNASCT**YP PVDKM*RVAPH	ninguna identidad
FRPPTRVLVSSVNMPDRMHPPIESLARTRLLGAV*LTRN RPMPQLGTPRR	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase (Sphingomonas gellani) 43.2%
TCLQIIYSMI*LRLNFYRRSPTALLDGHLPGEVRVI*QPIP GRTISPPVP	Uncharacterized protein (Thalassiosira oceanica) 42.1%
RI*L*RHSSSRKVRPAGMACL*FT*RRKIKDS*HQIAPVK CSPPTGITL*	ninguna identidad
PTLNGR*GKIGKDSGSLSTDLAGAGWSNKNGKCRVHEL DTRPFSFERTAL	solute carrier family 12-member 4 isoform X4 (Mesocricetus auratus) 42.9%
LNILTFRV*VVQSNYERCFIILFRPRMGSSGHRDRNTRTR ISDIE*RRCL	Uncharacterized protein (Branchiostoma floridae) 63.2%
ESAGVY*GPSPRTLRSSQVFRCNEDAFPTVYAWLQMSR YHMPFGMEEGPR	Zinc finger and AT-hook domain containing (Mastacembelus armatus) 41.2%
AIRRWNTQKHTRMPVPRVETISYSLRAPGLGFPSGNDTR *RRHA*YGFAG	2-isopropylmalate synthase (Mycobacteroides franklinii) 48.6%
PRGEGGSLKRRDASLLAV*IQYGSPDSGHTSARALRSEL SHNCLWSARHH	Uncharacterized protein (Micromonospora sp. MW-13) 40.0%
VSCRAF*VFNIVVKMRLAYISREDFMSHHQLGRVQS*RH CQV*SVSRAEN	ninguna identidad
PWI*VQATNQDGYGRL**YLFHACLVANN*HSKRPFEGS ESRSGKFRSFA	ninguna identidad
WEGADEVTMQQGDMKGWT*WLHM*FG*VGHRAYPIS AS*PAHPHWQGGTT	ninguna identidad
NEWVRHCQTTRGLGTLKFQNPTIGSCRTTILRVELPFLRS RAPARYTVVP	Uncharacterized protein (Anisakis simplex) 48.5%
GTHILKRSTSLTVSALTPRSSRGLGEDPSGCTEVQVDSCG APR*YVERDG	Asparaginase (Luteimicrobium xylanilyticum) 48.8%
FNLSANGFE*EKLESDL*EAKYNGGPHLSFEATSFWTLV LPEDLTAESGG	TIGR03557 family F420-dependent LLM class oxidoreductase (Glaciihabitans sp. INWT7) 55.6%
YRLVRTLVYWVSSG*DA*GPVRINRWLIAYGARTGPWL EIQQCRGCLRCS	DnaJ_C domain-containing protein (Perkinsus olseni) 51.7%

V*QYIIGTHD*SEPDALRSIYLH*NCSNTATGQVPFGQTII QGLF*APYS	ninguna identidad
RSK*VAWDGNMG*LSSPRSPNRSG*ADTLRPQASGRKTF KPSDLKKRCWP	RNA-directed DNA polymerase (Amphibalanus amphitrite) 51.7%
KV*WNRITSWPQIPDAPRLLSTITSARMEYRCIQVNDRK CSRMSKSENSK	Uncharacterized protein (Trichinella britovi) 60.0%
TCGTYVCIPRIL*LTDTCFRGF*HYFISRGCALYHRSVNW CKCASFRTKR	ninguna identidad
GRASLV*HLYRSSTS*WTNVIGLFNQPIPRDLPSSRKLIIY LCCRRGEHL	G/U mismatch-specific uracil-DNA glycosylase (Clostridiales bacterium) 41.7%
CAVHRPAI*PLRGGRRTDTLRISGKVTC*TDSSQSIPYVA RVIWC*RNNR	ninguna identidad
LHILHPVLPLTSSTYGRASAASVWGATVPSPATSQGQLR VSSKLP*SWR*	Uncharacterized protein (Pseudogymnoascus sp. VKM F-4…) 52.9%
ATDTLADPLPRLHSSCVASQRLLRHYISLKISTPAHVSGI RH*PPVHRDQ	Uncharacterized protein (Phytophthora fragariae) 42.1%
YLV*LSSQRERRNRTSSKGIGTVLSVRGSRFQALIAGLPC MRIYMKSCGC	ABC transporter permease (bacterium D16-50) 41.5%
MGNVAHSLDECSDLPPESRADGTALVEPEADILAVRDK SCIDTASMMRIR	OmpA family protein (Leucothrix pacifica) 45.5%
STQMTIRSSLQHSGGMTNNTHRPGEPERSAKDIWIVPRV PTLPVIIFAMS	Uncharacterized protein (Pterula gracilis) 50.0%
V*TSPTLKLALGEHKPKEGPAGYSSYSEFETLHGPLYSL* RTVQVLGILR	Uncharacterized protein (Orchesella cincta) 40.6%
LVDRFENHMLGHVA**YNRSWKRLNGRRLAVQ*GEMK SR**GDRVSLNY*	ninguna identidad
L*SSTNPP*HSGIETLDVRIIYLIRLAPPVLVDRSGGLKSG R*TPCAMLG	ninguna identidad
LPPKWKGSQRGTSCLVLGDHLMS*HWFCNFAGPSH*TA ID*MLPYSNKVA	ninguna identidad
LMRTYCGGCVVDNEGYPAILPALHI*GPRDF*V*RHSCG PCHSYGHRTNN	ninguna identidad
YICVKRILR*RGNTSVQ*QAY*GNPSCKLYYRILLKRSEP VSVGLPEGPY	ninguna identidad
M*IDSRRILLTYSCTRQREKKQLARPLVSFGMASDPQSA RRYSGKRYSQ*	J domain-containing protein (Penicillium solitum) 50.0%
SAGLIRQRGSSRE*YNQVASGLLRGLRTT**WVQRSPLQ AEAAIELPTIN	ninguna identidad
G**QELTKHDWLYGFGERLKLPRWSVYFILGWGRRRKS SYGLL*VWLAHP	Glutathionylspermidine synthase (Methylomonas sp. DH-1) 39.5%
ILVMARYLSTSFAHDAYNIPIDLAILVHQESSEASRHKV MVFTTNLFSRR	Uncharacterized protein (Corynebacterium sp. sy039) 40.6%
A*TGGIVQTL*ELNRA*GVCAILVSRPSYRVLRWLHYKW SSGDALSRARE	Uncharacterized protein (Halosimplex carlsbadense 2-9-1) 48.7%
RTHNRIILLTTHTLRIERIAQFQCFVIFSNIRTL*PIRVDNT METSATQA	ninguna identidad
TIPER*QSSQACRPAERMTIRIYFWPRLRTIQNLKSPLIDR TGSVATHEE	SPOR domain-containing protein (Cryobacterium tepidiphilum) 53.3%
HFCSLHAYGMLPANRMHRRRWRKNYKE*LKKD*SAST DLPLPS*R*AWRN	ninguna identidad
WAYVFGNQKRCLRSLHG*YGYPAGRAV*LCL*MQYLM SAF*RAKSLHTVK	ninguna identidad
D*TEIIG*RV*MFLE*VESRDGYPMT*HSRVRNI*RNHPKS PALTVVFPL	ninguna identidad
SSTAVKRVSGSALWYS*HAVGAMITLLWRLYAGTVRD VKFCPA*VMYVGV	Uncharacterized protein (Gemmataceae bacterium) 53.1%
SSVPRLKTGHSTGALHYLDIREYIPVYVS**TMQRRRQD *IEPLSKNLSL	ninguna identidad
DPA*TYRARRIRKLKSAFV*QPGEAQIRT*LRFNDSIGGV PVQRLVAGKV	ninguna identidad

STTERAALVTSSLRTIKYKRTPTRTGNGRRLTPKVDSRHTGHTFFPSQVC	DNA polymerase alpha subunit B (fungal sp. No.14919) 48.3%
*NHRNAAMLSYD*VKTTAVDYL*FNQYDMRLR*AERTHLNGLSEI*CWHA	ninguna identidad
MHPPVVLDFPFLSSAHRYLNREAILP*TRGASEYWCTPRIASEGIVMNLS	Uncharacterized protein (Phialemoniopsis curvata) 41.4%
GHVRSRQLAPCYHVLAKKHVLSWSS**LC*LSQDLTQAR*VGRRTKRPSN	Toll-like receptor 13 (Cyprinodon variegatus) 58.3%
EFIAGKSQRHLNKIYGT*QCSFSAGGILVRGTIGSITCHLVIDESRCACT	DNA protecting protein DprA (Sulfurisoma sediminicola) 51.5%
*VEGTFLPLGFSDCARVILKISSLNRWNISRVE*VMLRGI*DRAQGCRCR	ninguna identidad
*AKII*RPLL*GCIPVHTDISALLYFDSR*VHSQFDGEGGSCVSRRTGS*	ninguna identidad
QPATGRCLACPHVGYGKVRILGRPNTCESPRLLSAGKISSSYWLIRSPFI	Tripartite tricarboxylate transporter substrate binding protein (Betaproteobacteria bacterium) 59.1%
IIRSSVVSWNDRNNEESGYAPAKQGASLIRPVGFGIHPRRTVLHSDVPVA	SpoIID/LytB domain protein (Brockia lithotrophica) 42.1%
*PSGNPLESALACLSSAPHVPVRKFAEHSGRLRELCCDTSMRHFCGCGLG	Uncharacterized protein (Sphingomonas sp. Leaf34) 50.0%
LSISIPVSCASTWNLCRLVDVPLYNAKYTGSVVSCVGHSALFYSRPNALG	LRR receptor-like serine/threonine-protein kinase FLS2 (Cajanus cajan) 48.6%
*MRTS*LYRPAGKN*RQSAASAVNTSSWPRTPVLV**LGYTGSLTSKGNL	Uncharacterized protein (Ectocarpus sp. CCAP 1310/34) 54.2%
SGL*IYLWTRRNIHRHEGVLANL*RVASAIFTTRVYMYNDAHCA*RLSSN	ninguna identidad
QDQSFPVRALPELAIGKRCGCVRIELPVRHM**AAHINELVHAPY*TNIF	CN hydrolase domain-containing protein (Rhodotorula taiwanensis) 64.0%
EGSRWTRYAGIPHALCALMCSTGISTRTRA*LILFPAVTKIASPASCN*T	ninguna identidad
*RLDIVSAFICY*HQ*QKHTTWQLFPEAQLVIK*KDILFLRPLLALILRL	ninguna identidad
PRT*DNYS*SPRGDFNVKYRP*CLFLYGVWPRRDSRQSPNFLHPL*REAI	ninguna identidad
TPRVGHPSRQTIQGMKRHHPMYSVVVSQFPSHP*FLLPCKRCYVANREYI	Uncharacterized protein (Heterobasidion irregulare s…) 50.0%
FAPSFPRQTIDAANCHPALGRACAGVSQILERAW*GS*CPGCQNVQERFP	Putative RNA polymerase sigma factor (Azoarcus sp. (strain BH72)) 56.0%
KAVRDVGNR*RYIGYFFAIIVHSLYRRDHVSWSGASRQPVCTFRLDEKYF	Cbb3-type cytochrome c oxidase subunit 3 (Candidatus Afipia apatlaquen…) 41.4%
FLTCSVFQKCI*HRCAQTLFYRYPIAPDLLKTIQLEKRTRT*VGVFRSVKX	ninguna identidad
PWGRFPTGH*VDVGAYVVP*WSRYSRLPLCITLNVLFSNSDRSARTK*MG	ninguna identidad
L*RLGGRMRI*QSVNG*PLFLGNSLHESNDVVTPLR*VSTALR*SYGSIX	ninguna identidad
LMWCRS*YHIAYLQYAWALTPWRASMPYGVLIGPLLVRWST*YPVCRGWX	Uncharacterized protein (Sphingomonas sp.) 55.0%
KRTQMLADGHKITGSTLAWSDFALLRVSAVLLAQFFLFASMLSVSGDRPX	ABC-2 type transport system permease protein (Haladaptatus litoreus) 66.7%
YVGHSIPRLGGIDRRNT*GAAVEIGSLQ*YWGLFILRNSLCQVFSLRIGX	Bile acid:sodium symporter (Micromonospora sp. NRRL B-16…) 50.0%
GPHEDPQYSLSSVRPQHIKLYSKLHC*KISILQSQVRLVP*FLGEK*SHX	ninguna identidad
DVTTMLSLTEFFGINGKKALLLLAVTYGTKPRLQCIVMLILNQLVKKDPX	ninguna identidad
LMSQPPSYFSPPEPGVLGATPGPCSIRRFTRHAVHTARPDILA*RSVIP	Uncharacterized protein (Marchantia polymorpha) 71.4%
ANCGRQQRVRRKGTRGDRKEGV*L*PVATPTPPLSHLVRYDYSY*LGYP	Uncharacterized protein (Setaria italica.) 56.5%

KHMCRNQTPNGSVWKETMRRPGN*RKNS*RPYTQSFRS VRPLVR*RGIA	Uncharacterized protein (Auraticoccus monumenti) 52.2%
TAILTRLIPHLGSPKEPKEGYT*MTSRMSS*RAFRIYRRQS FAPYREDY	ninguna identidad
PPCKNFTHTHTP*F*A*IETPTIPRETPLR**GTGMSTGLV VAPCIRCQ	ninguna identidad
SLGRADSLPRSGDFAAGSGADIVHAFVTEFGLAPRTDLV GYF*H*SLMC	NADH_4Fe-4S domain-containing protein (Mycolicibacterium anyangense) 44.4%
PGVANGILTHESSGLGRA*RGNAMLSAPLQRMFVAVLE VSRLVP*QGIS	Uncharacterized protein (Salipiger profundus) 55.0%
YKQYTGSSFYLHPRCILLVCTRLISDR*GMLFSLGYQAG GHCHRTPYNT	Trace amine associated receptor 5 (Phasianus colchicus) 52.2%
LWATATEAL*VQWSPEGRALFLLDWGHTWWQ*RMAQ YNKLNGVIMNPVE	F-box WD repeat-containing 7-like (Micractinium conductrix) 60.9%
GGNSFGLPELRKVTHAQASGHNSSQRYLQWEWCSCRST HCIRHMCHIDT	Uncharacterized protein (Leishmania major) 44.7%
*ALHDLPRTPGACLAQPLCSS*TLYPIRRGPLC*ANCCGS PDSGVGNS*N	ninguna identidad
GFIAI*GRFIGVRRVLRTRRVVIPAQHTRLMNSLPRFRYR RL*L*EVCR	Probable membrane transporter protein (Roseomonas cervicalis ATCC 4…) 52.8%
RSGLTL*GRQNKFVAIL*ECGFGRRIKQHSMNGTDGTDK GRCST*ALND	ninguna identidad
KRLEQYLV*D*TGRRD*FCCPYVKVLLDFRVVTPSKNA GCRRPALKQVW	ninguna identidad
QSYGSGLMYSVIMYQNGSSQYC*WAQQNPL*IW*SIWP RSNFARRFTQG	DAGKa domain-containing protein (Gongylonema pulchrum) 52.4%
HVKAGAPLFMGDRRVGGQL*SFEVMRNLDLTTERPYLN *IDCAMTIPPD	ninguna identidad
VLEVCGSIK**GWSASSERYSYGRSDDFTQSVVLKSYYS VFGNFVTARK	Autotransporter domain-containing protein (Kluyvera genomosp. 3) 45.2%

Apéndice C

Tabla 3: Datos Estadísticos Relevantes Para el Análisis de Probabilidad. La última fila representa a las 42 secuencias con porcentaje de identidad igual a cero. X=valor de identidad obtenido, $\bar{X}$= la media del porcentaje de identidad

ID	$\boldsymbol{X}$	$\boldsymbol{X - \bar{X}}$	$\boldsymbol{(X - \bar{X})^2}$
1	0.615	0.3192	0.1018
2	0.455	0.1592	0.0253
3	0.488	0.1922	0.0369
4	0.750	0.4542	0.2062
5	0.520	0.2242	0.0502
6	0.577	0.2812	0.0790
7	0.432	0.1362	0.0185
8	0.421	0.1252	0.0156
9	0.429	0.1332	0.0177
10	0.632	0.3362	0.1130
11	0.412	0.1162	0.0135
12	0.486	0.1902	0.0361
13	0.400	0.1042	0.0108
14	0.485	0.1892	0.0357
15	0.488	0.1922	0.0369
16	0.556	0.2602	0.0677
17	0.517	0.2212	0.0489
18	0.517	0.2212	0.0489
19	0.600	0.3042	0.0925
20	0.417	0.1212	0.0146
21	0.529	0.2332	0.0543
22	0.421	0.1252	0.0156
23	0.415	0.1192	0.0142
24	0.455	0.1595	0.0253
25	0.500	0.2042	0.0416
26	0.406	0.1102	0.0123
27	0.500	0.2042	0.0416
28	0.395	0.0992	0.0098
29	0.406	0.1192	0.0142
30	0.487	0.1912	0.0365
31	0.533	0.2372	0.0562
32	0.531	0.2352	0.0553
33	0.483	0.1872	0.0350
34	0.414	0.1182	0.0139
35	0.583	0.2822	0.0824
36	0.515	0.2192	0.0480
37	0.591	0.2952	0.0871
38	0.421	0.1252	0.0156
39	0.500	0.2042	0.0416
40	0.486	0.1902	0.0361
41	0.582	0.2462	0.0606
42	0.640	0.3492	0.1219
43	0.500	0.2042	0.0416
44	0.560	0.2642	0.0698
45	0.550	0.2542	0.0646
46	0.667	0.3712	0.1377
47	0.500	0.2042	0.0416

48	0.714	0.4182	0.0174
49	0.562	0.2692	0.0724
50	0.522	0.2262	0.0511
51	0.444	0.1482	0.0219
52	0.550	0.2542	0.0646
53	0.522	0.2262	0.0511
54	0.609	0.3132	0.0980
55	0.447	0.1512	0.0228
56	0.528	0.2322	0.0539
57	0.524	0.2282	0.0520
58	0.452	0.1562	0.0243
59	0	-0.2958	0.0875