



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
DOCTORADO EN CIENCIAS BIOMÉDICAS
INSTITUTO DE INVESTIGACIONES BIOMÉDICAS

MODELO PROBABILÍSTICO DE EVOLUCIÓN DE PROTEÍNAS AUNADO A LA
EVOLUCIÓN DE LOS CÓDIGOS DEL TRNA

TESIS
QUE PARA OPTAR POR EL GRADO DE:
DOCTOR EN CIENCIAS

PRESENTA:
GABRIEL ZAMUDIO SOLÓRZANO

DIRECTOR DE TESIS
DR. MARCO ANTONIO JOSÉ VALENZUELA
INSTITUTO DE INVESTIGACIONES BIOMÉDICAS
COMITÉ TUTOR
DR. DANIEL PIÑERO DALMAU
INSTITUTO DE ECOLOGIA
DR. ARTURO CARLOS II BECERRA BRACHO
FACULTAD DE CIENCIAS

CIUDAD DE MÉXICO ENERO DE 2021



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

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

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

Dedicatorias

A mi familia y amigos, su apoyo incondicional durante mi vida me ha brindado la fuerza para siempre superarme y mejora en todos los aspectos de mi vida.

A la Universidad Nacional Autónoma de México, por darme la formación académica y profesional en tantas etapas de mi vida.

Agradecimientos

A mi director de tesis, el Dr. Marco Antonio Jos. Valenzuela, por ser pieza fundamental en mi formación y mi aprendizaje. Gracias por su atención y tiempo dedicado. Al Dr. Daniel Piñero Dalmau y al Dr. Arturo Carlos Il Becerra Bracho por formar mi comité tutorial y apoyarme durante todo mi trayecto en el doctorado.

Un agradecimiento al Mtro. Juan R. Bobadilla, por su asistencia técnica en computación durante el desarrollo de mi proyecto doctoral.

Un agradecimiento a Lucía Brito Ocampo y Martha Cariño Aguilar de la biblioteca del Instituto de Investigaciones Biomédicas UNAM, por proveerme el acceso a una colección de material bibliográfico completo y actualizado

Un agradecimiento al Grupo de Biología Teórica de Instituto de Investigaciones Biomédicas UNAM, por su apoyo durante la realización de mi doctorado.

Índice

Introducción	1
Objetivo	1
Hipótesis.....	2
El código genético	2
Modelo matemático del código genético	2
La unicidad del código genético	4
El código genético óptimo	4
Simetrías en el código genético estándar y otros códigos	5
Evolución de tRNAs	7
Elementos de identidad del tRNA	7
Elementos de identidad del tRNA en los tres dominios de la vida.	9
Evolución de proteínas	9
Modelo de evolución neutral	9
El efecto de la asimetría del código genético en la evolución de proteínas.	10
Conclusiones	12
Artículos fuera de la línea de la tesis doctoral	13
Historia antigua del centro de transferencia peptídica	13
Detección temprana de taquiarritmias inminentes	13
Conjunto mínimo de genes para un diagnóstico de carcinoma pulmonar de células escamosas	14
Referencias	15
Apéndices	19

Introducción

La información genética de los organismos está almacenada y codificada en las cadenas de nucleótidos de DNA (Ácido desoxirribonucleico) y RNA (Ácido ribonucleico). Los nucleótidos, a su vez, están formados por una molécula de azúcar (desoxirribosa para DNA y ribosa para RNA), un grupo fosfato y una base nitrogenada. Las bases nitrogenadas que forman el DNA y RNA son: adenina (A), guanina (G), citosina (C) y timina (T) en el caso de DNA, mientras que en RNA el uracilo (U) es usado en lugar de T. Los nucleótidos se dividen en dos grupos basándose en su estructura química: las purinas (R), y las pirimidinas (Y). Las purinas tienen dos anillos de carbono nitrógeno y las pirimidinas solo uno. Los nucleótidos A y G son purinas mientras que los nucleótidos C, T y U son pirimidinas.

La información del DNA es usada en procesos de replicación, producción de proteínas, procesos regulatorios, entre otros [1]. En el caso de la síntesis de proteínas, se requiere de la transcripción del DNA en una molécula de RNA mensajero (mRNA), el cual al estar en el ribosoma de la célula es traducido por moléculas de RNA de transferencia (tRNA). El tRNA es la molécula encargada de transferir el ordenamiento específico de tripletes de nucleótidos definido por el mRNA en proteínas por medio del ribosoma [1]. Antes de llegar al ribosoma, los tRNAs son precargados con su aminoácido específico por las enzimas aminoacil tRNA sintetasas (aaRSs), estas enzimas reconocen los 20 aminoácidos de forma específica y al mismo tiempo reconocen los distintos tRNAs que codifican para un mismo aminoácido [2]. De esta forma existen 20 aaRSs diferentes, una para cada aminoácido formando un código no degenerado [2]. Los tRNAs reconocen el mRNA por medio de un anticodón que es específico para cada codón, por lo tanto, existen varios tRNAs para un mismo aminoácido, al conjunto de tRNAs para un mismo aminoácido se le conoce como isoaceptores [3]. Las enzimas aaRSs se dividen en dos clases, clase I y clase II de acuerdo con la forma en que reconocen los tRNAs [2], las aaRSs clase I distorsionan la base terminal 3', mientras que la clase II no la distorsionan [4]. Se ha mostrado que las aaRSs no reconocen a los tRNAs modernos por medio del anticodón [5]. Por lo tanto el correcto reconocimiento de una aaRSs con un tRNA es por medio del código operacional, el cual se encuentra en la estructura del tRNA [6,7].

El código que relaciona tripletes de nucleótidos en el DNA con aminoácidos es el código genético estándar y está conformado por 64 posibles tripletes que codifican 20 aminoácidos y una señal de paro, este código se considera casi universal [8]. La relación de 64 codones a 20 aminoácidos y una señal de paro hace que el código genético sea degenerado, i.e, múltiples codones codifican para el mismo aminoácido, con excepción de los aminoácidos triptófano y metionina que son codificados por un codón cada uno. El tRNA contiene dos códigos en su estructura: el código de anticodones [9,10], el cual reconoce los codones del mRNA, y el código operacional [6,11], el cual reconoce el aminoácido correcto que debe cargar.

Objetivo

A partir del desarrollo de un modelo del código genético basado en el álgebra, analizar las características biológicas que le confieren su unicidad. Considerando el modelo evolutivo RNY del código genético se analizará su nivel de optimización. Comparar la estructura algebraica del código genético estándar a la de otros códigos mitocondriales. Derivar un modelo estocástico de evolución de proteínas y proponer una medida de neutralidad basada en dicho modelo. Analizar el efecto de la estructura del código genético en la medida de neutralidad. Usar la teoría de información para encontrar la relación entre el código de anticodones y el código estereoquímico de los tRNA. Detectar los sitios en la estructura del tRNA que determinan que estos sean aminoacilados correctamente.

Hipótesis

Las hipótesis consideradas durante el desarrollo de esta tesis son las siguientes:

- Si existe un conjunto finito de propiedades biológicas que le confieren sus características únicas entonces su nivel de optimización basado en dichas propiedades biológicas puede ser analizado a lo largo de su evolución.
- Si se desarrolla un modelo probabilidades de transición de aminoácidos basado en el código genético y la teoría de evolución neutral molecular entonces es posible calcular un control neutral de evolución que determine la frecuencia esperada de cada aminoácido en una proteína que sufre mutaciones al azar.
- Si en el proceso de identificación entre un tRNA y su aaRS correspondiente no se reconoce directamente el anticodón, entonces los tRNA cuentan con mecanismos en su secuencia que le permiten identificarse con la molécula aaRS correspondiente y su aminoacilación con el aminoácido correcto.

El código genético

Modelo matemático del código genético

El origen de la vida se calcula que ocurrió aproximadamente hace 3.7-3.8 mil millones de años [12] con genes que codificaban proteínas ribosomales y otras proteínas que estabilizaron el proceso de traducción genética [13]. Estos primeros organismos han sido llamados primer ancestro común universal (FUCA), por sus siglas en inglés [14]. El establecimiento de un código genético común para todos los organismos se llevó a cabo cuando el proto-organismo FUCA se madura [13], aproximadamente hace 2 mil millones de años [13]. El código genético primitivo estaba lejos del código genético actual que poseen los organismos conformado por 64 tripletes de nucleótidos o codones que codifican para 20 aminoácidos y una señal de paro. La propuesta más establecida del origen y evolución del código genético es la hecha por Manfred Eigen [15] basada en un código primitivo que considera únicamente tripletes de la forma RNY (purina-cualquiera-pirimidina) y se ha mostrado cómo es posible generar el código genético de 64 codones a partir de este subcódigo [16]. La extensión del código RNY se ha modelado algebraicamente usando teoría de grupos y campos algebraicos para representar los fenómenos biológicos de cambio de marco de lectura y mutaciones en la primera o tercera base de los tripletes RNY [16–18]. El fenómeno de cambio de marco de lectura se da cuando los tRNAs inician la traducción del mRNA a partir del segundo o tercer nucleótido y generar los codones de la forma NYR y YRN, a esta extensión del código RNY le nombramos código extendido 1. Las mutaciones puntuales se clasifican en transiciones, que no cambian el tipo químico del nucleótido y las transversiones, que son aquellas que cambian el tipo químico, es decir, sustituyen una purina por una pirimidina y viceversa; lo que permite generar codones de la forma RNR y YNY, a esta extensión del código RNY le nombramos código extendido 2. La unión del código extendido 1 y el código extendido 2 dan como resultado el código genético de 64 codones.

Se propuso un modelo matemático del código genético que pudiera ser usado para analizar de forma teórica diferentes propuestas sobre la evolución del código genético, así como las diferentes características que posee. El modelo matemático se basa en la teoría de grupos, en particular, la representación del grupo de Klein de 4 elementos (K_4) como mutaciones actuando sobre el conjunto de nucleótidos A, G, C, U para RNA representados con los vértices de un cuadrado (**Figura 1a**). El grupo algebraico K_4 es un grupo de orden 4, abeliano y con dos generadores. La asociación de los dos generadores de K_4 con los dos tipos de mutaciones biológicas de transición y transversión permiten transformar un nucleótido en otro por medio de la aplicación de los elementos de K_4 sobre el conjunto de nucleótidos. La extensión de la aplicación del grupo K_4 sobre tripletes de nucleótidos se da de forma natural al considerar tripletes ordenados de elementos de K_4 , lo que permite relacionar cualquier par de codones por medio un triplete de elementos

del grupo K_4 . Dado que los elementos de K_4 son auto-invertibles, i.e, son su propio inverso bajo la operación del grupo, el triplete $x = (x_1, x_2, x_3)$ de elementos de K_4 que relaciona los codones $a = (a_1, a_2, a_3)$ y $b = (b_1, b_2, b_3)$ por medio de la ecuación $xa = (x_1a_1, x_2a_2, x_3a_3) = (b_1, b_2, b_3) = b$ también cumple la ecuación $xb = a$. La representación de los nucleótidos en un cuadrado resulta en que el conjunto de 64 codones se puede representar como los vértices de un hipercubo de seis dimensiones (**Figura 1b**). Dos codones $a = (a_1, a_2, a_3)$ y $b = (b_1, b_2, b_3)$ tendrán un arista que los una en el hipercubo si $a_i = b_i$ en dos posiciones y en la posición donde $a_j \neq b_j$ se cumpla que $xa_j = b_j$ donde x es un generador del grupo K_4 .

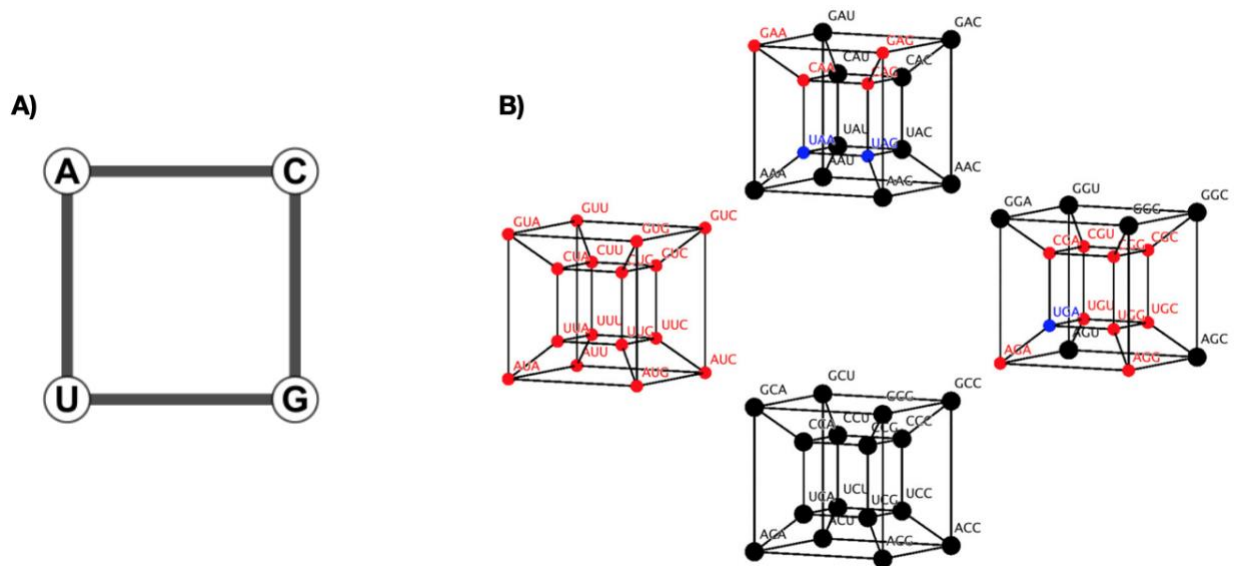


Figura 1 A) Representación de los 4 nucleótidos de RNA como los vértices de un cuadrado. **B)** Representación de los 64 codones del código genético estándar en un hipercubo de 6 dimensiones, codones asociados a aminoácidos de clase I en rojo, codones asociados a aminoácidos de clase II en negro, codones asociados a la señal de paro en azul.

Esta representación algebraica de los codones permitió comparar diferentes modelos de evolución del código genético. En particular, se analizó el modelo antes descrito que considera un subcódigo primitivo RNY y la generación del código completo por medio de cambios en el marco de lectura y mutaciones puntuales en la primera y tercera base de los codones, también se analizó el modelo Delarue [19] y el modelo Rodin-Ohno [20–22]. El modelo de Delarue está basado en la hipótesis de un código inicial ambiguo NNN que obtuvo especificidad por medio de cinco particiones binarias de forma jerárquica que actuaron en base a las particiones de los nucleótidos en el grupo de pirimidinas y purinas. El modelo Rodin-Ohno se basa en la observación de que la tabla del código genético puede ser dividida de forma casi simétrica al considerar las dos clases de aaRSs que se encargan de reconocer los tRNAs para cargarlos con su aminoácido específico y que las dos clases de aaRSs pudiesen ser descendientes de un mismo RNA de doble cadena ancestral. Estos dos hechos dan como resultado que los grupos de codones NAN, NGN, NCN, NUN sean la partición natural del código genético ya que permitirían mantener la simetría dada por las aaRSs.

El modelo algebraico del código genético permitió dar una función explícita que transformara los grupos de codones RNY, YNY, RNR y YNR del modelo RNY en los grupos del modelo Rodin-Ohno, lo cual mostró la equivalencia de estos dos modelos evolutivos del código genético. También se describe un automorfismo del hipercubo de codones que intercambia las dos clases de aaRSs, reflejando así la

simetría de las dos clases de aaRSs. Por otro lado, el modelo algebraico permite representar las divisiones binarias del modelo Delarue por medio de cocientes algebraicos de la forma N/K_4 donde N es el conjunto de nucleótidos. El desarrollo detallado del modelo algebraico, la equivalencia entre el modelo RNY con el modelo Rodin-Ohno y la representación teórica del modelo Delarue se encuentran en el apéndice 1.

La unicidad del código genético

La representación del código genético en un hipercubo de seis dimensiones aunado con el grupo K_4 para representar los cambios de nucleótidos en los codones nos da un modelo teórico con el cual podemos resolver diferentes hipótesis acerca de la evolución con código genético. Una de las primeras cuestiones surge a partir del planteamiento de Francis Crick de la hipótesis del accidente congelado [23], la cual propone que el código genético es universal ya que cualquier cambio al que fuere expuesto sería letal y por lo tanto el cambio sería suprimido. Esta hipótesis de unicidad del código genético en principio no considera un camino evolutivo de éste.

Usando el modelo RNY de evolución del código genético, que considera como código inicial el código RNY y su ampliación por medio de los cambios de marco de lectura junto con transversiones en la primera y tercera base de los codones, junto con el hipercubo de codones sometemos a prueba la hipótesis de si existe un arreglo de codones equivalente en el hipercubo que se pueda obtener a partir del modelo evolutivo RNY y las propiedades biológicas del código genético. Las propiedades biológicas que consideramos son la degeneración del código genético, las propiedades de wobble que ocurren principalmente en la tercera base de los codones, la no degeneración en la asociación de los aaRSs con los aminoácidos y su distinción en dos clases, la consideración del aminoácido glicina como primer aminoácido codificado por el código genético [24,25]. La equivalencia de códigos genéticos la definimos como cuando existe una asociación diferente entre los codones y los aminoácidos que mantenga todas las propiedades biológicas que consideramos. Esta asociación diferente se traduce matemáticamente en que exista un automorfismo en el hipercubo que transforme un código en otro. Con este conjunto de parámetros biológicos y la representación en forma de hipercubo del código genético se mostró que la asociación entre codones y aminoácidos del código genético estándar es la única que puede resultar al considerar el modelo evolutivo RNY. La unicidad del código genético al considerar un modelo evolutivo representa la selección de un código de un total de 2.81×10^4 posibles códigos. Los detalles del desarrollo de esta sección se encuentran en el apéndice 2.

El código genético óptimo

La cualidad de óptimo del código genético y su capacidad para tolerar mutaciones que cambien los aminoácidos o la estructura química de los aminoácidos durante el proceso de traducción se ha analizado ampliamente considerando diferentes metodologías y se ha encontrado que el código genético estándar es sub-óptimo en análisis de desempeño [26–32]. A partir del modelo del hipercubo se sometió a prueba la hipótesis de que el código genético estándar es sub-óptimo en un contexto de códigos genético al azar, pero es óptimo cuando se considera el proceso evolutivo dado por el modelo RNY. Para esto se considera el subcódigo RNY como código inicial y se consideran los códigos extendidos 1 y 2 como puntos intermedios en la evolución del código genético de 64 codones. Se diseñaron 15,000 códigos divididos en tres grupos, cada grupo con diferentes niveles de aleatoriedad para ser comparados con el código genético estándar. Como primer nivel de aleatoriedad se consideran códigos con la asignación de los 20 aminoácidos y la señal de paro completamente al azar con los 64 codones. En un segundo nivel de aleatoriedad se consideran códigos donde los aminoácidos mantienen su codonicidad, es decir, el número de codones que los codifican. En un tercer nivel de aleatoriedad se consideran códigos donde se mantiene la codonicidad de cada aminoácido y también se mantiene la propiedad de wobble en la tercera base. El tercer nivel de aleatoriedad es el que genera códigos más similares al código genético estándar.

La asociación de los codones con su respectivo aminoácido es la partición natural del código genético y por lo tanto se pueden definir los conceptos de clases de equivalencia y cocientes algebraicos. A una gráfica de vértices y aristas se le puede aplicar el cociente algebraico con respecto a una clase de equivalencia de forma análoga a la definición usual sobre conjuntos. El cociente de una gráfica con respecto a una clase de equivalencia nos dará como resultado una gráfica cociente en la cual los vértices estarán dados por las clases de equivalencia, dos vértices en la gráfica cociente estarán unidos por un arista si existen vértices en la gráfica original de cada clase de equivalencia que estén unidos por un arista. Considerando el modelo del código genético en un hipercono como una gráfica y la partición del código genético natural se calcula la gráfica cociente, la cual es llamada gráfica fenotípica [10,33]. La gráfica fenotípica describe las relaciones que existen entre los codones del código genético, ya que dos aminoácidos de la gráfica fenotípica estarán unidos si estos aminoácidos tienen codones que los codifiquen que estén a una mutación puntual de distancia. Las medidas de conectividad de la gráfica fenotípica nos reflejan el ordenamiento de los codones en el modelo del hipercono, si la gráfica fenotípica presenta más aristas, entonces los aminoácidos están más cercanos entre sí y por lo tanto se requieren menos mutaciones puntuales para cambiar la codificación de un triplete de un aminoácido a otro. La medida de centralidad de una gráfica que nos habla sobre el grado de conectividad general de una gráfica es la conectividad algebraica [34,35], la cual está dada por el segundo eigenvalor más pequeño diferente de cero de la matriz laplaciana de una gráfica. A medida que la conectividad algebraica de una red aumenta, su conectividad aumenta. Si consideramos un código genético en el que no se cumpliera la propiedad de wobble en la tercera base y por lo tanto los aminoácidos estuviesen distribuidos de forma homogénea en el conjunto de codones, la gráfica fenotípica de este código tendría una conectividad algebraica más alta que el código genético estándar, ya que la propiedad de wobble agrupa los codones que codifican para un mismo aminoácido.

Al calcular la conectividad algebraica de los tres grupos de códigos aleatorios en las diferentes etapas evolutivas del código genético se encontró que para el código RNY y el extendido 1 presentan el valor mínimo de conectividad algebraica, mientras que por otro lado existieron códigos aleatorios que al considerarse los 64 codones o la restricción dada por el código extendido 2 tienen valores más pequeños de conectividad que el código genético estándar. Sin embargo, los códigos aleatorios que al tomar en cuenta los 64 codones presentaron valores de conectividad algebraica más pequeños que el código genético estándar presentaron alteraciones en el subcódigo RNY, por lo tanto, estos códigos tendrían un origen evolutivo distinto del código genético estándar. Estos resultados confirman lo encontrado por otros autores al afirmar que en general el código genético estándar no es el óptimo que podría existir y presenta evidencias de que si consideramos un origen a partir de un código RNY si es el código genético óptimo. Los detalles del desarrollo de esta sección se encuentran en el apéndice 3.

Simetrías en el código genético estándar y otros códigos

La organización del código genético dada por el wobble en la tercera posición de los codones le da una organización en bloques al código. En la tabla del código genético, los codones para un mismo aminoácido, en general, están agrupados en bloques donde los primeros dos nucleótidos son iguales y en la tercera base hay diferencias, para los aminoácidos con dos codones o di-codónicos, los codones están relacionados por una transición en la tercera base. La organización en bloques da una aparente simetría en la estructura organizacional del código genético. A partir de la representación en un hipercono es posible describir matemáticamente las posibles simetrías presentes en el código genético estándar [33]. El código genético estándar en casi universal [8], algunos organismos poseen variaciones en las que los aminoácidos selenocisteína y pirrolisina están presentes y son principalmente codificados por codones asociados a la señal de paro en el código genético estándar [36,37]. También, el código genético de las mitocondrias posee variantes más considerables y sus códigos genéticos se han considerado como evoluciones del código genético estándar [38]. El modelo del hipercono nos sirve para comparar las estructuras de código genético estándar y los códigos mitocondriales. Es de notar que el hipercono del código genético surge a partir de la representación de los nucleótidos como vértices de un cuadrado, sin embargo, existen tres posibles asociaciones de los nucleótidos con los vértices que no son equivalentes (**Figura 2**). Es posible usar también la representación de los nucleótidos en un cuadrado con diagonales

que representa la posibilidad de cambiar un nucleótido en cualquier otro por medio de una única mutación (**Figura 2**).

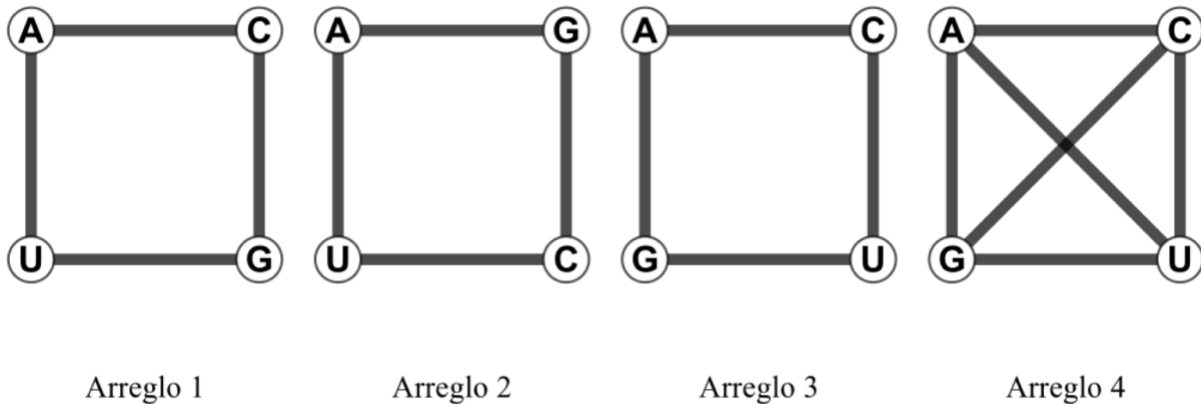


Figura 2 Las tres diferentes representaciones de los nucleótidos como vértices de un cuadrado y la representación en un cuadrado con las diagonales.

Estas cuatro representaciones de los nucleótidos generan hipercubos que son diferentes entre sí, donde todos son igualmente válidos, por lo tanto, es necesario el análisis de todos para describir la estructura organizacional de los códigos genéticos y como ésta cambia en el modelo evolutivo RNY, tanto en el código genético estándar como en los códigos mitocondriales. Para lo cual se analizó la forma en que el modelo del hipercubo se genera a partir del subcódigo RNY, se describe el grupo de automorfismos del hipercubo que mantiene invariante la asociación de codones con aminoácidos y los automorfismos presentes en las respectivas gráficas fenotípicas. Se encontró que los automorfismos en el hipercubo de codones que mantienen invariantes los aminoácidos, son de orden pequeño en el subcódigo RNY y se pierden en el camino al código genético estándar. Por otro lado, el código genético mitocondrial de diferentes grupos de organismos presenta más simetrías que el código estándar (**Tabla 3**). Las gráficas fenotípicas presentaron simetrías en muy pocos casos. Los detalles del desarrollo de esta sección se encuentran en el apéndice 4.

Aminoácido	Código Genético Estándar		Código Mitocondrial Invertebrados		Código Mitocondrial Vertebrados		Código Mitocondrial Levadura	
Ala	GCA	GCC	GCA	GCC	GCA	GCC	GCA	GCC
	GCG	GCU	GCG	GCU	GCG	GCU	GCG	GCU
Arg	CGA	CGC	CGA	CGC	CGA	CGC	CGA	CGC
	CGG	CGU	CGG	CGU	CGG	CGU	CGG	CGU
	AGA	AGG					AGA	AGG
Asn	AAC	AAU	AAC	AAU	AAC	AAU	AAC	AAU
Asp	GAC	GAU	GAC	GAU	GAC	GAU	GAC	GAU
Cys	UGC	UGU	UGC	UGU	UGC	UGU	UGC	UGU
Gln	CAA	CAG	CAA	CAG	CAA	CAG	CAA	CAG
Glu	GAA	GAG	GAA	GAG	GAA	GAG	GAA	GAG
Gly	GGA	GGC	GGA	GGC	GGA	GGC	GGA	GGC
	GGG	GGU	GGG	GGU	GGG	GGU	GGG	GGU
His	CAC	CAU	CAC	CAU	CAC	CAU	CAC	CAU
Ile	AUA	AUC	AUC	AUU	AUC	AUU	AUC	AUU
	AUU							
Leu	UUA	UUG	UUA	UUG	UUA	UUG	UUA	UUG
	CUA	CUC	CUA	CUC	CUA	CUC		
	CUG	CUU	CUG	CUU	CUG	CUU		
Lys	AAA	AAG	AAA	AAG	AAA	AAG	AAA	AAG
Met	AUG		AUG	AUA	AUG	AUA	AUG	AUA
Phe	UUC	UUU	UUC	UUU	UUC	UUU	UUC	UUU
Pro	CCA	CCC	CCA	CCC	CCA	CCC	CCA	CCC
	CCG	CCU	CCG	CCU	CCG	CCU	CCG	CCU
Ser	UCA	UCC	UCA	UCC	UCA	UCC	UCA	UCC
	UCG	UCU	UCG	UCU	UCG	UCU	UCG	UCU
	AGC	AGU	AGC	AGU	AGC	AGU	AGC	AGU
					AGA	AGG		
Stop	UAA	UAG	UAA	UAG	UAA	UAG	UAA	UAG
	UGA		AGA	AGG				
Thr	ACA	ACC	ACA	ACC	ACA	ACC	ACA	ACC
	ACG	ACU	ACG	ACU	ACG	ACU	ACG	ACU
							CUA	CUC
						CUG	CUU	
Trp	UGG		UGG	UGA	UGG	UGA	UGG	UGA
Tyr	UAC	UAU	UAC	UAU	UAC	UAU	UAC	UAU
Val	GUA	GUC	GUA	GUC	GUA	GUC	GUA	GUC
	GUG	GUU	GUG	GUU	GUG	GUU	GUG	GUU

Tabla 3 El código genético estándar y tres códigos genéticos mitocondriales diferentes.

Evolución de tRNAs

Elementos de identidad del tRNA

La correcta implementación del código genético esta mediada por un extenso conjunto de interacciones biológicas dentro de la célula. En el centro de este sistema se encuentran los RNA de transferencia (tRNA), moléculas encargadas de la implementación del proceso de traducción a través del reconocimiento de los codones del RNA mensajero (mRNA) por medio del ribosoma [39]. Un tRNA es una molécula de aproximadamente 76 nucleótidos [40] con algunas variaciones. Los tRNAs también cuentan con una sección CCA terminal que es añadida después de la transcripción [41] y es constante en todos los tRNAs ya que es el sitio de carga de los aminoácidos. Dentro de su estructura contiene dos códigos, el código de anticodones y el código operacional [6]. El código de anticodones se encarga del reconocimiento de los codones del mRNA y el código operacional se encarga del reconocimiento de las aaRSs que están

previamente cargadas con un aminoácido [42]. El código de anticodones se encuentra en la zona de anticodones, en los nucleótidos 34, 35 y 36, y el código operacional se encuentra en la región aceptora (**Figura 3a**) [43]. Se ha hipotetizado que estos códigos han co-evolucionado desde el origen del tRNA hace 3.5 billones de años [44]. Cada tRNA puede reconocer un aminoácido y un conjunto de codones que lo codifiquen por medio del efecto wobble. El conjunto de tRNAs diferentes que reconocen un mismo aminoácido se le conoce como isoaceptores. Con el uso de la teoría de información se evaluó la hipótesis de que cada conjunto isoaceptor reconoce la misma aaRS entonces deben tener el mismo código operacional. La medida conocida como variación de información (VI) mide, dadas dos variables aleatorias X y Y , la distancia entre ellas, i.e, la cantidad de información requerida para determinar completamente una variable aleatoria a partir de la otra. La variación de información se calcula con la fórmula $VI(X, Y) = H(X) + H(Y) - 2I(X, Y)$, donde $H(X)$ es la entropía de la variable aleatoria X e $I(X, Y)$ es la información mutua entre las variables aleatorias X y Y .

Se utilizó una base de datos curada [45] que contiene las secuencias de tRNAs para los 20 aminoácidos canónicos. Se seleccionaron los tRNA con una longitud de 68 nucleótidos al descartar la región variable y el CCA terminal y se removieron secuencias duplicadas. En total se analizaron 13, 093 secuencias para los 20 isoaceptores. Para cada isoaceptor se consideró cada posición como una variable aleatoria y se calculó la variación de información entre todas las posiciones de cada isoaceptor. Obtener como resultado una variación de información con valor de 0 significa que las dos posiciones del isoaceptor están completamente relacionadas entre sí, es decir, es posible determinar que nucleótido hay una posición a partir de la otra. Por lo tanto, los sitios encargados de implementar el código operacional estarán a una distancia en información de 0 con las posiciones del anticodón y se formará un agrupamiento.

Para cada isoaceptor se encontraron distintas posiciones que están a una distancia en información de 0 de las posiciones del anticodón, además se encontraron otros agrupamientos de posiciones a distancia 0 que no contienen al anticodón (**Figura 3b**). A los distintos agrupamientos de bases a distancia 0 les nombramos elementos de identidad ya que se encontró que varían entre los isoaceptores. En particular, las posiciones que se agrupan con el anticodón se propusieron como implementadoras del código operacional. Los detalles del desarrollo de esta sección se encuentran en el apéndice 5.

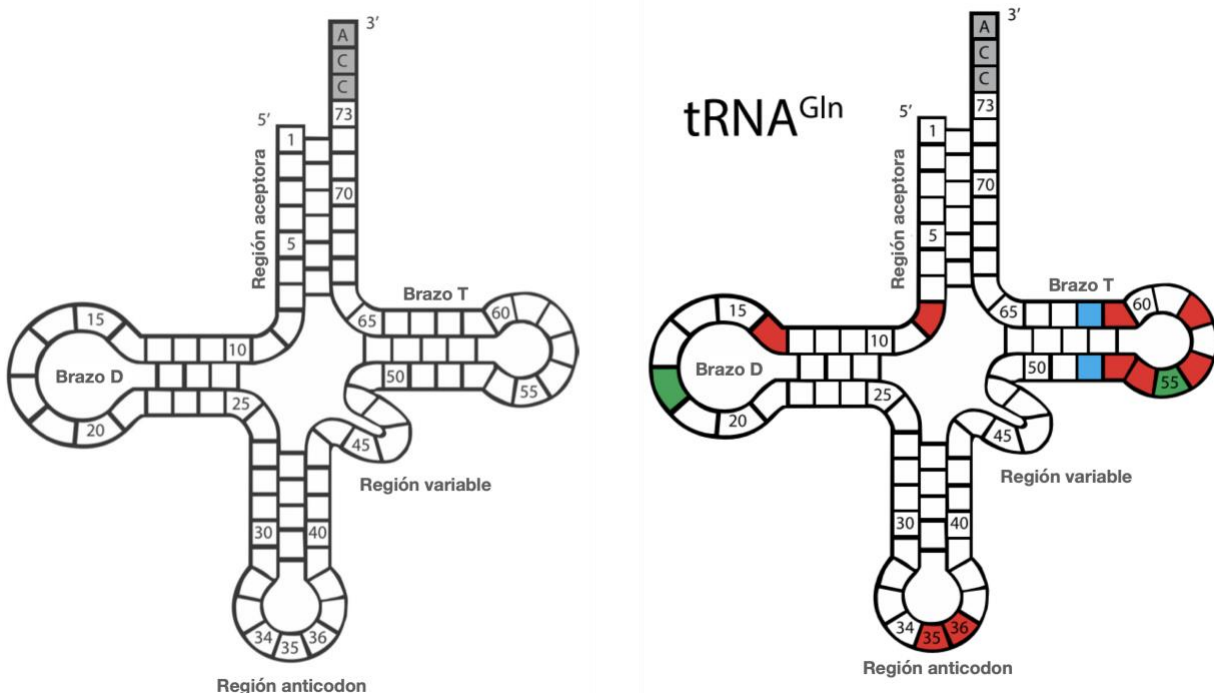


Figura 3 A) Estructura secundaria del tRNA, las bases correspondientes al anticodón están marcadas de la 34 a la 36. B) Los tres agrupamientos de posiciones a distancia de información 0 presentes en el tRNA de glutamina en verde, azul y rojo. Las posiciones asociadas con el anticodón están en rojo.

Elementos de identidad del tRNA en los tres dominios de la vida.

Al encontrar elementos de identidad en los 20 isoaceptores de tRNA nos planteamos la hipótesis de si estos conjuntos de posiciones en el tRNA son los mismos en los tres dominios de la vida distinguidos por Carl Woese, bacteria, archaea y eukarya [46,47]. Para esto se consideraron dos bases de datos [45,48], se hizo la división de los isoaceptores en los tres dominios de la vida, se usaron las secuencias con una longitud de 68 nucleótidos al retirar la región variable y el CCA terminal. A continuación, se aplicó la misma metodología de variación de información para el cálculo de agrupamientos de posiciones. Para cada isoaceptor en cada uno de los dominios se calculó el valor mínimo de variación de información que formaba agrupamientos bien definidos. Se encontró que los isoaceptores de archaea tienen un mayor número de agrupamientos, seguido de eukarya y por último el dominio de bacteria con un menor número de posiciones de identidad en los isoaceptores. Estos resultados son consistentes con lo encontrado en la literatura sobre la forma en que se encuentran los genes de tRNA en cada dominio. En archaea se han encontrado tRNAs funcionales que solo están formados por la región aceptora y la de anticodones [49], por lo tanto, es necesario que sus tRNAs contengan un mayor número de elementos de identidad para la correcta implementación del código genético. En general se puede observar que existen diferencias en los elementos de identidad de los tres dominios y sus coincidencias son las reportadas en un trabajo previo. Los detalles del desarrollo de esta sección se encuentran en el apéndice 6.

Evolución de proteínas

Modelo de evolución neutral

Las gráficas fenotípicas generadas a partir de la representación del código genético en un hipercono de seis dimensiones se pueden extender a gráficas dirigidas con pesos en las aristas. Dado un aminoácido, el conjunto de codones que lo codifica está enlazado a otros codones en modelo del hipercono, dicho conjunto de adyacencias representa todos los posibles cambios de ese aminoácido con una mutación puntual en alguno de los codones. Por lo tanto, es posible contar las sinónimas y no sinónimas, i.e, que cambian o no el aminoácido codificado, de un aminoácido. Si asignamos pesos uniformes a los cambios puntuales entre codones, se calcula la suma de pesos que intercambian los codones de un aminoácido en los codones de otro aminoácido y se obtiene así una gráfica fenotípica con pesos en las aristas. Al normalizar los pesos de los vértices de salida de cada aminoácido para que sumen uno se obtiene la probabilidad de que un aminoácido cambie en otro. Con estas probabilidades podemos definir un proceso estocástico de Markov con tiempo discreto y con el conjunto de aminoácidos como estados. El proceso estocástico definido es irreducible y dado que tiene un conjunto de estados finito, entonces todos los estados son recurrentes, en particular positivos recurrentes. Dado que los estados son positivos recurrentes entonces existe una distribución estacionaria del proceso estocástico y es única. Esta distribución estacionaria está asociada a la gráfica fenotípica, a la representación en seis dimensiones del código genético y a la representación de los nucleótidos como vértices de un cuadrado. Dado que existen tres ordenamientos para los nucleótidos como vértices de un cuadrado, se pueden derivar las respectivas tres matrices de probabilidades de transición, las cuales al promediarse dan lugar a un proceso estocástico que considera igualmente los tres ordenamientos. Este promedio de procesos tiene un sesgo hacia mantener transiciones ya que en dos de los tres ordenamientos de los nucleótidos las transiciones están representadas por una arista del cuadrado. La distribución estacionaria del proceso estocástico promedio la proponemos como control de evolución neutral. La propiedad de neutralidad surge a partir de la asignación de pesos uniformes en las mutaciones puntuales que modifican codones, por lo tanto, el cambio de un codón en otro es completamente por azar. Este proceso fue propuesto previamente por Kimura en la teoría de evolución neutral [50], la cual establece que la mayoría de los cambios a nivel molecular están guiados por cambios al azar aunados a la deriva genética y no por procesos de selección [51]. La mencionada teoría fue extendida posteriormente a considerar cambios que son casi neutrales o que tienen bajos niveles de presiones de selección [52]. La distribución estacionaria como control neutral nos expresa la probabilidad de observar un aminoácido en una proteína hipotética que solo está sujeta a mutaciones

neutrales. Por lo tanto, al comparar el control con una proteína específica podemos evaluar si un aminoácido en específico ha sufrido presiones de selección positivas, negativas o sufre cambios neutrales. Como ejemplo de la aplicación del modelo se comparó el control de evolución neutral con las proteínas Histona 4, Citocromo C, Citocromo C oxidasa, Fibrinógeno alfa, β – Hemoglobina y la proteína de shock térmico 90. Para cada proteína se consideraron 100 secuencias diferentes descargadas de la base de datos UniProt [53]. El conjunto de secuencias de cada proteína fueron alineadas con el software MUSCLE [54], a partir del alineamiento múltiple se hizo una matriz que contabilizara los cambios de codones de cada proteína y se derivó una matriz de probabilidades de cambios de aminoácidos. A partir de la matriz de probabilidades de cambio de aminoácidos se derivó su distribución estacionaria, la cual fue comparada con el control neutral y analizada. También se consideró la comparación del control neutral con otros modelos de evolución ampliamente usados como por ejemplo el BLOSUM62 [55], se derivó la distribución estacionaria del modelo de evolución neutral del código mitocondrial de vertebrados y la distribución estacionaria derivada de considerar los nucleótidos en un cuadrado con las diagonales que representa el modelo de sustituciones propuesto por Jukes - Cantor [56] (**Figura 5**). Se observó que el modelo de evolución Jukes – Cantor da lugar a una distribución estacionaria esencialmente igual al control neutral propuesto con muy ligeras variaciones en los valores numéricos, por otro lado, se resalta que al considerar un código genético diferente como lo es el código mitocondrial de vertebrados se obtiene un control neutral diferente. Dado que el modelo de evolución BLOSUM62 surge a partir de considerar una gran cantidad de proteínas y derivar un modelo que aproxime su camino evolutivo, es natural que este modelo lleve implícito las presiones evolutivas de las proteínas y por lo tanto su distribución estacionaria difiera de la del control neutral. El control de evolución neutral propuesto puede ser usado para analizar la evolución de cualquier proteína ya que la única hipótesis que se considera en su derivación es el código genético. Los detalles del desarrollo de esta sección se encuentran en el apéndice 7.

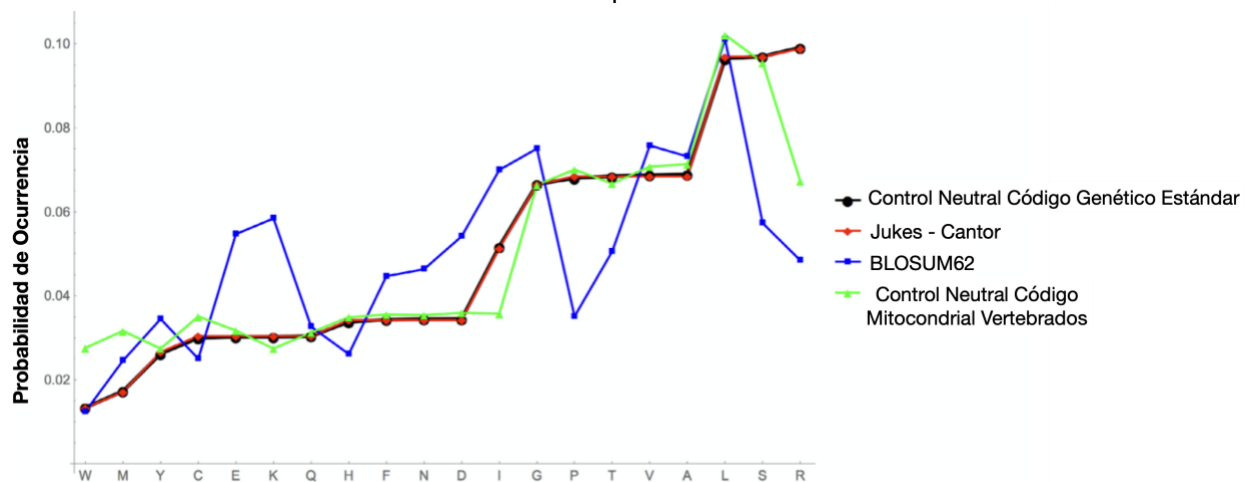


Figura 5. Distribución estacionaria del modelo neutral de evolución (negro). Distribución estacionaria del modelo Jukes-Cantor (Rojo). Distribución estacionaria del modelo BLOSUM62 (azul). Distribución estacionaria del modelo neutral considerando el código genético mitocondrial de vertebrados (verde).

El efecto de la asimetría del código genético en la evolución de proteínas

El modelo de evolución neutral tiene como hipótesis la estructura del código genético para su derivación. El control de evolución neutral a la vez cumple que los aminoácidos con la misma codonicidad tienen probabilidades similares de ocurrir, esto se observa en los aminoácidos de dos y cuatro codones. En la representación usual del código genético en forma de tabla, se pueden resaltar algunas propiedades, por ejemplo, la propiedad de wobble y en que la degeneración en la tercera base de los codones genera una estructura de bloques en la distribución de los aminoácidos. Anteriormente se describieron los grupos de automorfismos existentes en las gráficas fenotípicas del código genético, las cuales están compuestas

principalmente por los grupos \mathbb{Z}_2 y S_3 en los diferentes arreglos de nucleótidos en un cuadrado. Las órbitas dadas por los automorfismos relacionan aminoácidos que tienen la misma codonicidad. Por lo tanto, se hizo el siguiente planteamiento sobre la igualdad en los valores del control neutral: Si la propiedad de que los aminoácidos con la misma codonicidad presenten los mismos valores surge a partir de la estructura de bloques del código genético o bien del arreglo particular de los bloques en el código genético. Para esto se diseñaron tres códigos genéticos sintéticos que mantuvieran la degeneración de cada aminoácido, el wobble en la tercera posición, pero que su distribución fuese diferente (**Figura 6**). En estos tres códigos la señal de paro fue considerada como otra señal codificada por el código genético para ser asociada a diferentes codones.

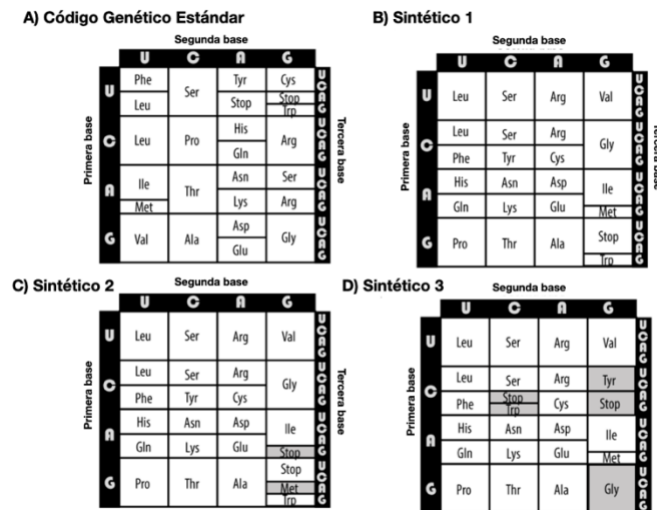


Figura 6. A) Tabla del código genético estándar resaltando su estructura en bloques. B) Código sintético 1 con aparente mayor simetría en su organización. C) Código sintético 2 con las diferencias respecto al código sintético 1 resaltadas en gris. D) Código sintético 3 con las diferencias respecto al código sintético 1 resaltadas en gris.

Para los tres códigos sintéticos se calcularon sus gráficas fenotípicas y sus respectivos grupos de automorfismos, también se calculó el control neutral asociado a ellos y se compararon con el código genético estándar (**Figura 7**). Se encontró que los controles neutrales presentaban diferentes valores y los aminoácidos relacionados a través de los automorfismos tienen los mismos valores. Los códigos sintéticos fueron diseñados para tener mejor organización de los aminoácidos con la misma codonicidad. Eso se refleja en que, en particular para el código sintético 1, el control neutral es equiprobable para los aminoácidos con la misma codonicidad. Por lo tanto, la estructura del código genético estándar da como resultado que incluso bajo condiciones de mutaciones neutrales, algunos aminoácidos tengan más ocurrencia que otros a pesar de tener la misma codonicidad. En las gráficas asimétricas, la selección de un aminoácido no está condicionada a la selección de cualquier otro aminoácido. La asimetría observada es una propiedad que deja intacta la universalidad del código genético. El código genético termina su influencia inmediatamente después de la aminoacilación de cada tRNA. La asimetría de las gráficas fenotípicas ha permitido la sorprendente diversidad de los organismos. Los detalles del desarrollo de esta sección se encuentran en el apéndice 8.

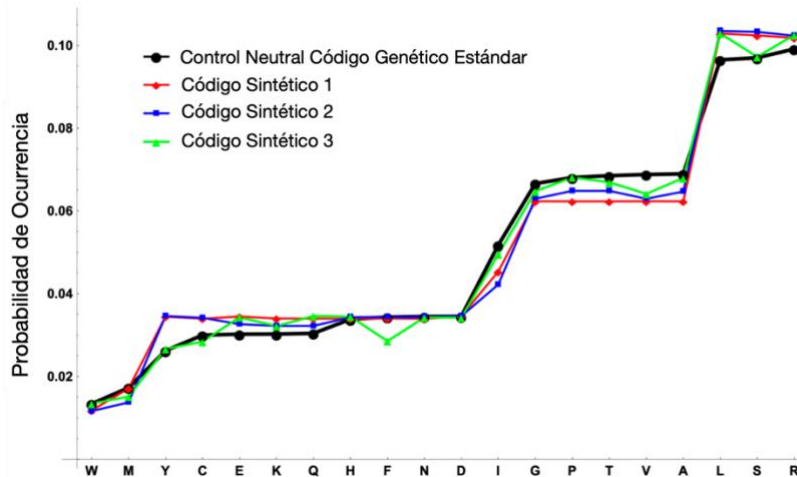


Figura 7. Distribución estacionaria del modelo neutral de evolución considerando el código genético estándar (negro) y la distribución estacionaria de los códigos sintéticos.

Conclusiones

Desde el desciframiento del código genético [23], una de las propuestas más aceptadas, aunada a la de Manfred Eigen, sobre la evolución del código genético es la hecha por Crick que describe un accidente congelado [23]. Por lo tanto, esta propuesta es poco descriptiva acerca de la evolución del código genético. El modelo matemático aquí planteado del código genético, es la plataforma para analizar modelos de evolución del código genético desde un enfoque matemático y teórico. La representación geométrica del código genético en 6 dimensiones y su transformación en redes fenotípicas nos permitió analizar diferentes aspectos de él. Se analizó la cuestión de si el código genético es óptimo y se describió el escenario en el que sí resulta ser óptimo, considerando como óptimo el que las mutaciones puntuales que sufre en codón puedan ser interpretadas como mutaciones sinónimas en el proceso de traducción. También, con base en la representación en 6 dimensiones, se describieron las características biológicas que le dan el carácter de único y se comparó con otros códigos genéticos que existen en la célula.

Las moléculas de tRNA son centrales en la implementación del código genético. Con base en un análisis de teoría de información, se describió el código operacional, el cual es el mecanismo encargado del correcto reconocimiento entre un tRNA y su correspondiente aaRS. Se mostró cómo este código operacional es diferente para cada aminoácido y también es diferente en los tres dominios de la vida. Por otro lado, a partir del modelo en 6 dimensiones del código genético se derivó un modelo neutral de evolución que es tan universal como el código genético estándar y, por lo tanto, es aplicable a los tres dominios de la vida. El control de evolución neutral nos describe la proporción de aminoácidos que tendría una proteína hipotética que evolucionara completamente al azar y no estuviese sujeta a ningún tipo de presión evolutiva. Las gráficas fenotípicas y el modelo de evolución neutral nos permitieron analizar la forma en que las variaciones en la estructura del código genético estándar darían lugar a variaciones en el control neutral de evolución. Los resultados de estos trabajos, principalmente los relacionados al código operacional del tRNA, deben ser estudiados más profundamente en el contexto experimental para poder ser validados y usados en áreas de la ciencia aplicada.

Artículos fuera de la línea de la tesis doctoral

Durante el desarrollo de esta tesis se realizaron tres trabajos en colaboración, cuyo tema está fuera de la línea de investigación del proyecto doctoral y se describen a continuación:

Historia antigua del centro de transferencia peptídica

El centro de transferencia peptídica (PTC por sus siglas en inglés) es el centro catalítico del ribosoma, se encuentra en la subunidad grande del ribosoma y es responsable de la unión de aminoácidos junto con el alargamiento de péptidos durante la síntesis de proteínas. El PTC se considera crucial para el entendimiento del origen de la vida al ser un catalizador de una relación mutualista entre los ácidos nucleicos y péptidos, permitiendo así el origen de la vida [57–59]. Se han propuesto diversas teorías sobre el origen y evolución del PTC [60,61].

Para este trabajo se consideró la propuesta que considera la formación del PTC a partir de la concatenación de cinco tRNAs [62]. Se utilizó la metodología de variación de información utilizada en los trabajos anteriores sobre tRNA [3,63] para analizar la conservación de sitios en el PTC. Se utilizaron un total de 1434 secuencias de la molécula 23S ribosomal de la base de datos GenBank [64], a las cuales se les extrajo la sub-secuencia de 179 nucleótidos asociados al PTC. Las secuencias de PTC fueron alineadas con el software ClustalW [65], posteriormente se calculó la variación de información entre cada posición del alineamiento. Se consideraron los diferentes grupos de sitios que comparten una variación de información de 0. A partir del modelo del PTC por concatenación de tRNAs se hizo un modelado molecular para derivar su estructura secundaria con el software UGENE [66] y terciaria con la plataforma ModeRNA [67]. La simulación de la estructura secundaria y terciaria se comparó con la estructura actual del *T. thermophilus*. Al identificar los diferentes tRNAs concatenados en la estructura secundaria de PTC y mapear los grupos de posiciones con variación de información de 0, se encontró que los diferentes grupos de posiciones se encuentran compartidos por dos tRNAs, marcando así un orden en la concatenación. También se presentó un grupo grande de posiciones relacionadas entre sí, el cual proponemos que puede estar relacionado con el plegamiento general del PTC. La estructura secundaria del PTC formada por la concatenación de tRNAs presenta tres sitios formando pares Watson-Crick dando uniones que no están presentes en el PTC moderno de *T. thermophilus*. Estos resultados refuerzan la propuesta del origen del PTC por medio de la concatenación de tRNAs. Los detalles del desarrollo de esta sección se encuentran en el apéndice 9.

Detección temprana de taquiarritmias inminentes

Los desfibriladores implantados ofrecen una ventaja sin precedente para los pacientes con la enfermedad de taquiarritmia ventricular inminente. Esta es una enfermedad que se caracteriza por episodios de taquiarritmia no predecibles que pueden conducir a la muerte cardiaca súbita del que la padece. La teoría de redes ofrece la posibilidad de analizar interacciones de cualquier tipo y se han convertido en una manera de analizar sistemas con dinámica compleja [68]. La terminología de “alerta temprana” no es, hasta ahora, parte de la comunidad cardiológica dado que las enfermedades cardiacas son la acumulación de muchos factores.

Existen algunos métodos para transformar series de tiempo en redes de interacción [69], los cuales han sido modificados y refinados para diferentes aplicaciones [70,71]. En este trabajo proponemos un nuevo método paramétrico para la transformación de series de tiempo en redes de interacción, las “ ε -regular graphs” (gráficas ε -regulares). Estas se construyen a partir de la previa definición de un parámetro ε que puede tomar el valor de cualquier número mayor o igual a cero. Una vez definido el valor del parámetro ε , los vértices de la gráfica ε -regular corresponden a los tiempos de la serie de tiempo. Las aristas de la red están definidas de la siguiente forma, dos vértices t_n, t_m de la serie de tiempo correspondientes a los tiempos n y m respectivamente estarán unidos en la red si se cumple que $|t_n - t_m| \leq \varepsilon$. Esta metodología se aplicó al análisis de series de tiempo correspondiente a los intervalos R-R de 81 pacientes

diagnosticados con taquiarritmia ventricular y con desfibriladores implantados. Los datos se extrajeron de la base de datos PhysioNet [72] y contienen una serie de intervalos R-R de 81 pacientes. Para cada paciente hay una serie control en la que el sujeto no sufre de ninguna alteración cardiaca y una serie de los últimos minutos previos a una taquiarritmia ventricular en la que la última medición es previa a la activación de dispositivo desfibrilador implantado. También se analizaron series de intervalos R-R de pacientes sin ninguna enfermedad cardiaca diagnosticada. Las series de tiempo se pre-procesaron para que tengan la misma cantidad de mediciones a partir de la última medición para incluir el momento de la activación del desfibrilador. Usando el método de ventanas deslizantes, se consideraron ventanas de 60 mediciones con un deslizamiento de una medición. Para cada ventana se calcularon las gráficas ε -regulares con un valor del parámetro $\varepsilon = 0.04$. Posteriormente se calculó la medida de centralidad de grado promedio de la red en cada ventana. Al comparar la serie de tiempo del grado promedio de cada ventana de los datos control y datos con episodio de taquiarritmia inminente se encontró un cambio en la dinámica de la serie de tiempo asociada al episodio de taquiarritmia 8:20 minutos antes de que ocurriera la taquiarritmia. También se encontró una gran diferencia entre la serie de tiempo del grado promedio de los individuos sin afecciones cardiacas. La metodología de gráficas ε -regulares puede dar lugar a la detección de alertas tempranas en taquiarritmia ventricular que pueden ayudar a salvar la vida de pacientes. Los resultados derivados de este análisis requieren estudios posteriores para poder ser eventualmente aplicados. Los detalles del desarrollo de esta sección se encuentran en el apéndice 10.

Conjunto mínimo de genes para un diagnóstico de carcinoma pulmonar de células escamosas

El cáncer pulmonar de células escamosas es uno de los tipos de cáncer pulmonar más común. En este tipo de cáncer el diagnóstico más usual se logra en los estados más avanzados de la enfermedad, lo que lo hace difícil de tratar. Las metodologías basadas en objetivos moleculares se han enfocado en otros tipos de cáncer hasta ahora [73–75]. En este trabajo nos planteamos el objetivo de determinar un conjunto mínimo de genes que puedan ser considerados como firma molecular de las diferentes etapas del desarrollo tumoral, para así ampliar las opciones disponibles de diseños experimentales y objetivos moleculares para el desarrollo de tratamientos específicos en cada etapa de este carcinoma. Se usaron los datos de expresión genética de 122 pacientes, en los que se incluyen los 7 estados clasificados del desarrollo tumoral y pacientes sin diagnóstico de carcinoma como control [76]. Estos datos de expresión genética se componen de los valores de expresión genética de 41,067 genes. Se entrenó un modelo de aprendizaje automático basado en una regresión logística usando como parámetros los datos de expresión de los 41,067 genes, a partir de ahí se diseñó una reducción de parámetros a partir de los coeficientes de cada parámetro en el modelo de la regresión logística. Los valores de los coeficientes del modelo se normalizaron y se consideraron los coeficientes que estuvieran a dos desviaciones estándar de la media. Este proceso da lugar a solo considerar los genes que son más importantes para el modelo de regresión logística. Se desarrolló un segundo modelo de regresión logística considerando solamente los genes encontrados en la primera reducción de parámetros y se procedió a hacer una segunda reducción de parámetros. Este proceso iterado dio lugar a la identificación de 15 genes específicos. Al entrenar un modelo de regresión logística usando como parámetros del modelo únicamente el conjunto de 15 genes identificados se llegó a un modelo capaz de identificar los diferentes estados del desarrollo del carcinoma con un índice de Jaccard de 0.92. Este modelo presentó clasificaciones erróneas en las fases finales del desarrollo tumoral. A continuación, se procedió a identificar los 15 genes en una red de interacción proteína-proteína, en donde se encontró la función molecular de estos genes. Por otro lado, al comparar con la literatura disponible se encontró que, de los genes identificados, algunos de ellos se han empezado a utilizar como objetivo de tratamientos moleculares, mientras que los otros forman parte de nuestra propuesta como genes a considerar en estudios experimentales. Los detalles del desarrollo de esta sección se encuentran en el apéndice 11.

Referencias

1. Alberts, B.; Johnson, A.; Lewis, J.; Morgan, D.; Raff, M.; Roberts, K.; Walter, P. *Molecular Biology of the Cell*; 2017;
2. Eriani, G.; Delarue, M.; Poch, O.; Gangloff, J.; Moras, D. Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. *Lett. to Nat.* **1990**, *347*, 203–206.
3. Zamudio, G.S.; Palacios-Pérez, M.; José, M. V. Information theory unveils the evolution of tRNA identity elements in the three domains of life. *Theory Biosci.* **2020**, *139*, 77–85.
4. Carter, C.W.; Wills, P.R. Hierarchical groove discrimination by Class I and II aminoacyl-tRNA synthetases reveals a palimpsest of the operational RNA code in the tRNA acceptor-stem bases. *Nucleic Acids Res.* **2018**, *46*, 9667–9683.
5. Ribas de Pouplana, L.; Schimmel, P. Operational RNA Code for Amino Acids in Relation to Genetic Code in Evolution. *J. Biol. Chem.* **2001**, *276*, 6881–6884.
6. De Duve, C. The second genetic code. *Nature* **1988**, *333*, 117–118.
7. Giegé, R.; Sissler, M.; Florentz, C. Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Res.* **1998**, *26*, 5017–5035.
8. José, M. V.; Zamudio, G.S.; Morgado, E.R. A unified model of the standard genetic code. *R. Soc. Open Sci.* **2017**, *4*, 160908.
9. Guimarães, R.C.; Moreira, C.H.C.; de Farias, S.T. A self-referential model for the formation of the genetic code. *Theory Biosci.* **2008**, *127*, 249–270.
10. José, M. V.; Morgado, E.R.; Guimarães, R.C.; Zamudio, G.S.; de Farias, S.T.; Bobadilla, J.R.; Sosa, D. Three-Dimensional Algebraic Models of the tRNA Code and 12 Graphs for Representing the Amino Acids. *Life (Basel, Switzerland)* **2014**, *4*, 341–73.
11. Schimmel, P.; Giege, R.; Moras, D.; Yokoyama, S. An operational RNA code for amino acids and possible relationship to genetic code. *Proc. Natl. Acad. Sci. U. S. A.* 1993, *90*, 8763–8768.
12. Joyce, G.F. The antiquity of RNA-based evolution. *Nature* 2002.
13. Prosdocimi, F.; de Farias, S.T. From FUCA To LUCA : A Theoretical Analysis on the Common Descent of Gene Families. *Acta Sci. Microbiol.* **2020**, *3*, 1–9.
14. Prosdocimi, F.; José, M. V.; de Farias, S.T. The First Universal Common Ancestor (FUCA) as the Earliest Ancestor of LUCA's (Last UCA) Lineage. In *Evolution, Origin of Life, Concepts and Methods*; Pontarotti, P., Ed.; Springer International Publishing: Cham, 2019; pp. 43–54 ISBN 978-3-030-30363-1.
15. Eigen, M.; Schuster, P. The Hypercycle - A principle of natural self-organization Part B: The abstract hypercycle. *Naturwissenschaften* **1978**, *65*, 7–41.
16. José, M. V.; Morgado, E.R.; Govezensky, T. An extended RNA code and its relationship to the standard genetic code: an algebraic and geometrical approach. *Bull. Math. Biol.* **2007**, *69*, 215–43.
17. José, M. V.; Morgado, E.R.; Sanchez, R.; Govezensky, T. The 24 possible algebraic representations of the standard genetic code in six or in three dimensions. *Adv. Stud. Biol.* **2012**, *4*, 119–52.
18. José, M. V.; Morgado, E.R.; Govezensky, T. Genetic hotels for the standard genetic code: evolutionary analysis based upon novel three-dimensional algebraic models. *Bull. Math. Biol.* **2011**, *73*, 1443–76.
19. Delarue, M. An asymmetric underlying rule in the assignment of codons: possible clue to a quick early evolution of the genetic code via successive binary choices. *RNA* **2007**, *13*, 161–169.
20. Rodin, S.N.; Rodin, A.S. On the origin of the genetic code: signatures of its primordial complementarity in tRNAs and aminoacyl-tRNA synthetases. *Heredity (Edinb.)* **2008**, *100*, 341–

355.

21. Rodin, S.N.; Ohno, S. Two types of aminoacyl-trna synthetases could be originally encoded by complementary strands of the same nucleic ACID. *Orig. Life Evol. Biosph.* **1995**, *25*, 565–589.
22. Rodin, S.N.; Rodin, A.S. Partitioning of aminoacyl-tRNA synthetases in two classes could have been encoded in a strand-symmetric RNA world. *DNA Cell Biol.* **2006**, *25*, 617–626.
23. Crick, F.H.C. The origin of the genetic code. *J. Mol. Biol.* **1968**, *38*, 367–379.
24. Bernhardt, H.S.; Patrick, W.M. Genetic code evolution started with the incorporation of glycine, followed by other small hydrophilic amino acids. *J. Mol. Evol.* **2014**, *78*, 307–309.
25. Tamura, K. Beyond the Frozen Accident: Glycine Assignment in the Genetic Code. *J. Mol. Evol.* **2015**, *81*, 69–71.
26. Alff-Steinberger, C. The genetic code and error transmission. *Proc. Natl. Acad. Sci. U. S. A.* **1969**, *64*, 584–591.
27. Ardell, D.H.; Sella, G. No accident: genetic codes freeze in error-correcting patterns of the standard genetic code. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **2002**, *357*, 1625–42.
28. Novozhilov, A.S.; Wolf, Y.I.; Koonin, E. V. Evolution of the genetic code: partial optimization of a random code for robustness to translation error in a rugged fitness landscape. *Biol. Direct* **2007**, *2*, 24.
29. Wong, J.T.-F. Role of minimization of chemical distances between amino acids in the evolution of the genetic code. *Proc. Natl. Acad. Sci. U. S. A.* **1980**, *77*, 1083–1086.
30. Haig, D.; Hurst, L.D. A Quantitative Measure of Error Minimization in the Genetic-Code. *J. Mol. Evol.* **1991**, *33*, 412–417.
31. Freeland, S.J.; Knight, R.D.; Landweber, L.F.; Hurst, L.D. Early Fixation of an Optimal Genetic Code. *Mol. Biol. Evol.* **2000**, *17*, 511–518.
32. Koonin, E. V.; Novozhilov, A.S. Origin and Evolution of the Universal Genetic Code. *Annu. Rev. Genet.* **2017**, *51*, 45–62.
33. José, M. V.; Zamudio, G.S.; Palacios-Pérez, M.; Bobadilla, J.R.; de Farias, S.T. Symmetrical and Thermodynamic Properties of Phenotypic Graphs of Amino Acids Encoded by the Primeval RNY Code. *Orig. Life Evol. Biosph.* **2015**, *45*, 77–83.
34. de Abreu, N.M.M. Old and new results on algebraic connectivity of graphs. *Linear Algebra Appl.* **2007**, *423*, 53–73.
35. Newman, M.E.J. *Networks : an introduction*; Oxford University Press, 2010; ISBN 9780199206650.
36. Srinivasan, G.; James, C.M.; Krzycki, J.A. Pyrrolysine encoded by UAG in archaea: Charging of a UAG-decoding specialized tRNA. *Science (80-)*. **2002**, *296*, 1459–1462.
37. Zinoni, F.; Birkmann, A.; Stadtman, T.C.; Bock, A. Nucleotide sequence and expression of the selenocysteine-containing polypeptide of formate dehydrogenase (formate-hydrogen-lyase-linked) from *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.* **1986**, *83*, 4650–4654.
38. Polacek, N.; Mankin, A.S. The ribosomal peptidyl transferase center: Structure, function, evolution, inhibition. *Crit. Rev. Biochem. Mol. Biol.* **2005**, *40*, 285–311.
39. Cusack, S. Aminoacyl-tRNA synthetases. *Curr. Opin. Struct. Biol.* **1997**, *7*, 881–889.
40. Altman, R.B. Probabilistic structure calculations: a three-dimensional tRNA structure from sequence correlation data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1993**, *1*, 12–20.
41. Hou, Y.-M. CCA addition to tRNA: Implications for tRNA quality control. *IUBMB Life* **2010**, *62*, 251–260.
42. Arnez, J.G.; Moras, D. Structural and functional considerations of the aminoacylation reaction. *Trends Biochem. Sci.* **1997**, *22*, 211–6.

43. Hou, Y.-M.; Schimmel, P. A simple structural feature is a major determinant of the identity of a transfer RNA. *Nature* **1988**, *333*, 140–145.
44. Tamura, K. Origins and Early Evolution of the tRNA Molecule. *Life* **2015**, *5*, 1687–1699.
45. Abe, T.; Inokuchi, H.; Yamada, Y.; Muto, A.; Iwasaki, Y.; Ikemura, T. TRNADB-CE: tRNA gene database well-timed in the era of big sequence data. *Front. Genet.* 2014, *5*, 114.
46. Woese, C.R.; Kandler, O.; Wheelis, M.L. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U. S. A.* **1990**, *87*, 4576–4579.
47. Woese, C.R.; Fox, G.E. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci. U. S. A.* **1977**, *74*, 5088–5090.
48. Jühling, F.; Mörl, M.; Hartmann, R.K.; Sprinzl, M.; Stadler, P.F.; Pütz, J. tRNADB 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.* **2009**, *37*, D159-62.
49. Fujishima, K.; Kanai, A. tRNA gene diversity in the three domains of life. *Front. Genet.* **2014**, *5*, 142.
50. Kimura, M. The neutral theory of molecular evolution. **1979**, *241*, 94–104.
51. Kimura, M. Recent development of the neutral theory viewed from the Wrightian tradition of theoretical population genetics. *Proc. Natl. Acad. Sci.* **1991**, *88*, 5969–5973.
52. Ohta, T. Slightly Deleterious Mutant Substitutions in Evolution. *Nature* **1973**, *246*, 96–98.
53. The UniProt Consortium UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **2017**, *45*, D158–D169.
54. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797.
55. Henikoff, S.; Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **1992**, *89*, 10915–10919.
56. Jukes, T.H.; Cantor, C.R. Evolution of Protein Molecules. In *Mammalian Protein Metabolism*; Elsevier, 1969; pp. 21–132 ISBN 978-1-4832-3211-9.
57. Lanier, K.A.; Petrov, A.S.; Williams, L.D. The Central Symbiosis of Molecular Biology: Molecules in Mutualism. *J. Mol. Evol.* **2017**, *85*, 8–13.
58. Vitas, M.; Dobovišek, A. In the Beginning was a Mutualism - On the Origin of Translation. *Orig. Life Evol. Biosph.* **2018**, *48*, 223–243.
59. de Farias, S.T.; Prosdociimi, F. *A emergência dos sistemas biológicos: uma visão molecular da origem da vida*; 1st ed.; ArtcomCiencia, 2019; ISBN 978-65-900624-1-3.
60. Agmon, I.; Bashan, A.; Zarivach, R.; Yonath, A. Symmetry at the active site of the ribosome: Structural and functional implications. *Biol. Chem.* 2005, *386*, 833–844.
61. Belousoff, M.J.; Davidovich, C.; Zimmerman, E.; Caspi, Y.; Wekselman, I.; Rozenszajn, L.; Shapira, T.; Sade-Falk, O.; Taha, L.; Bashan, A.; et al. Ancient machinery embedded in the contemporary ribosome. *Biochem. Soc. Trans.* 2010, *38*, 422–427.
62. de Farias, S.T.; Rêgo, T.G.; José, M. V. Origin and evolution of the Peptidyl Transferase Center from proto-tRNAs. *FEBS Open Bio* **2014**, *4*, 175–178.
63. Zamudio, G.S.; José, M. V. Identity Elements of tRNA as Derived from Information Analysis. *Orig. Life Evol. Biosph.* **2018**, *48*, 73–81.
64. Sayers, E.W.; Cavanaugh, M.; Clark, K.; Ostell, J.; Pruitt, K.D.; Karsch-Mizrachi, I. GenBank. *Nucleic Acids Res.* **2019**, *48*, D84–D86.
65. Thompson, J.D.; Gibson, T.J.; Higgins, D.G. Multiple Sequence Alignment Using ClustalW and ClustalX. *Curr. Protoc. Bioinforma.* **2003**, *00*, 2.3.1-2.3.22.
66. Okonechnikov, K.; Golosova, O.; Fursov, M.; team, the U. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics* **2012**, *28*, 1166–1167.

67. Rother, M.; Rother, K.; Puton, T.; Bujnicki, J.M. ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic Acids Res.* **2011**, *39*, 4007–4022.
68. Albert, R.; Barabasi, A.L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **2002**, *74*, 47–97.
69. Lacasa, L.; Luque, B.; Ballesteros, F.; Luque, J.; Nuño, J.C. From time series to complex networks: The visibility graph. *Proc. Natl. Acad. Sci.* **2008**, *105*, 4972–4975.
70. Gonçalves, B.A.; Carpi, L.; Rosso, O.A.; Ravetti, M.G. Time series characterization via horizontal visibility graph and Information Theory. *Phys. A Stat. Mech. its Appl.* **2016**, *464*, 93–102.
71. Bezsudnov, I. V.; Snarskii, A.A. From the time series to the complex networks: The parametric natural visibility graph. *Phys. A Stat. Mech. its Appl.* **2014**, *414*, 53–60.
72. Goldberger, A.L.; Amaral, L.A.N.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.-K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation* **2000**, *101*.
73. Gandara, D.R.; Hammerman, P.S.; Sos, M.L.; Lara, P.N.; Hirsch, F.R. Squamous cell lung cancer: From tumor genomics to cancer therapeutics. *Clin. Cancer Res.* **2015**, *21*, 2236–2243.
74. Singh, A.P.; Adrianzen Herrera, D.; Zhang, Y.; Perez-Soler, R.; Cheng, H. Mouse models in squamous cell lung cancer: impact for drug discovery. *Expert Opin. Drug Discov.* **2018**, *13*, 347–358.
75. Hashemi-Sadraei, N.; Hanna, N. Targeting FGFR in Squamous Cell Carcinoma of the Lung. *Target. Oncol.* **2017**, *12*, 741–755.
76. Sean, D.; Meltzer, P.S. GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **2007**, *23*, 1846–1847.

Apéndice 1

Research



Cite this article: José MV, Zamudio GS, Morgado ER. 2017 A unified model of the standard genetic code. *R. Soc. open sci.* 4: 160908.
<http://dx.doi.org/10.1098/rsos.160908>

Received: 14 November 2016

Accepted: 30 January 2017

Subject Category:

Genetics

Subject Areas:

bioinformatics/evolution/theoretical biology

Keywords:

standard genetic code, symmetry groups, aminoacyl-tRNA synthetases, group actions, automorphisms, polar requirement

Author for correspondence:

Marco V. José

e-mail: marcojose@biomedicas.unam.mx

A unified model of the standard genetic code

Marco V. José¹, Gabriel S. Zamudio¹ and

Eberto R. Morgado²

¹Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, Mexico D.F. 04510, Mexico

²Facultad de Matemática, Física y Computación, Universidad Central 'Marta Abreu' de Las Villas, Santa Clara, Cuba

MVJ, 0000-0001-8497-6681; GSZ, 0000-0003-4486-9843

The Rodin–Ohno (RO) and the Delarue models divide the table of the genetic code into two classes of aminoacyl-tRNA synthetases (aaRSs I and II) with recognition from the minor or major groove sides of the tRNA acceptor stem, respectively. These models are asymmetric but they are biologically meaningful. On the other hand, the standard genetic code (SGC) can be derived from the primeval RNY code (R stands for purines, Y for pyrimidines and N any of them). In this work, the RO-model is derived by means of group actions, namely, symmetries represented by automorphisms, assuming that the SGC originated from a primeval RNY code. It turns out that the RO-model is symmetric in a six-dimensional (6D) hypercube. Conversely, using the same automorphisms, we show that the RO-model can lead to the SGC. In addition, the asymmetric Delarue model becomes symmetric by means of quotient group operations. We formulate isometric functions that convert the class aaRS I into the class aaRS II and vice versa. We show that the four polar requirement categories display a symmetrical arrangement in our 6D hypercube. Altogether these results cannot be attained, neither in two nor in three dimensions. We discuss the present unified 6D algebraic model, which is compatible with both the SGC (based upon the primeval RNY code) and the RO-model.

1. Introduction

The insight that all organisms on Earth are related by common descent [1] is a remarkable scientific achievement. Indeed, the Last Universal Common Ancestor seemed to obey already the standard genetic code (SGC), which is *nearly universal*. The problem of the origin and evolution of the SGC is a fundamental challenge in biology. After the decipherment of the SGC [2], there have been several proposals that account for both the origin and evolution of the genetic code [3–11]. There seems to be a consensus that the SGC conserves vestiges of earlier codes, to wit, the operational [12,13] and anticodon codes [14,15]. The amino acid specific aminoacylation of tRNAs (operational

code) is localized in the acceptor stem of the tRNAs and is recognized by the corresponding aminoacyl-tRNA synthetases (aaRSs) [12,13]. Indeed, most living organisms still contain relics of these primeval codes, which are a palimpsest over which the evolving codes were later additions in order to arrive at the frozen SGC [16,17]. In fact, the primeval RNY code was already frozen [18].

The SGC is written in an alphabet of four letters (C, A, U, G), grouped into words three letters long, called triplets or codons. In general, and in most textbooks, the genetic code is represented in a two-dimensional (2D) table arranged in such a way that it is possible to readily find any amino acid from the three letters, written in the 5' to 3' direction of the codon [4,19,20]. Each of the 64 codons specifies one of the 20 amino acids or else serves as a punctuation mark signalling the end of a message. The standard table of codon assignments derives from the obvious representation of the triplet code as a $4 \times 4 \times 4$ cube. Three-dimensional (3D) algebraic models using a Galois Field of four elements GF(4) [21,22] or Lie algebras [22] have also been formulated. More revealing representations have been attained using the six-dimensional (6D) hypercube [23,24] of the 64 codons of the SGC. Observing that 64 is equal not only to 4^3 but also to 2^6 , the codon table can be organized as a 6D hypercube or 6D vector space $(\mathbb{Z}_2)^6$ over the binary field $\mathbb{Z}_2 = \{0, 1\}$ [24]. The phenotypic graphs of amino acids have been obtained from the topology of the SGC [15]. Additionally, circular representations of the SGC have been proposed [25–27]. Given 64 codons and 20 amino acids plus a punctuation mark, there are $21^{64} \approx 4 \times 10^{84}$ possible genetic codes. The result that only one in every million random alternative codes is more efficient than the SGC [28] implies that there could be approximately 4×10^{78} genetic codes as efficient as the SGC. This calculation does not offer deeper insights concerning the origin and structure of the SGC, particularly the frozen accident. Francis Crick [4] argued that the SGC need not be special at all; it could be nothing more than a 'frozen accident'. Yet as we show in this article, there are indeed several features that are special about the SGC: firstly, it can be partitioned *exactly* into two classes of aaRRs in six dimensions; secondly, it displays symmetry groups when the polar requirement (PR) is used; and thirdly, the SGC can be broken down into a product of simpler groups reflecting the pattern of degeneracy observed, and the salient fact that evolution did not erase its own evolutionary footsteps.

The search for symmetries in the SGC has been made by examining the tRNA [29,30] and aaRSs [3,6–8], using phylogenetic methods [31,32]. Less popular have been algebraic models seeking to unveil hidden symmetries of the SGC [33,34]. For example, the SGC has been theoretically derived from a primeval RNY (R means purines, Y pyrimidines, and N any of them) genetic code [9] under a model of sequential symmetry breakings [16,21,35]. Universal vestiges of these evolutionary steps were found in current genomes of both Eubacteria and Archaea [35]. The SGC is implemented via the tRNAs that bind each codon with its anticodon. These molecules define the genetic code, by linking the specific amino acids and tRNAs with the corresponding anticodons [7]. The tRNA molecule itself displays two codes, the operational code and the anticodon code. Typically, two genetic codes are considered, to wit, the 'classic' code represented in tRNA by an anticodon for reading codons in mRNA, and the other is the 'second' [12] operational RNA code [13,36] mapped mainly to the acceptor for appropriate aminoacylation at its 3' terminus. In addition, there are also two separate codes, embedded in the tRNA anticodon and acceptor-stem bases that correspond, respectively, to amino acid size and hydrophobicity [37,38]. These coding elements evolved separately and independently [38]. The earlier appearance of an acceptor-stem code, before the emergence of the universal genetic code [13] is supported experimentally by (i) the reciprocal biochemistry of minihelix acylation by full-length synthetases [39] and (ii) the acylation of full-length tRNAs by truncated synthetases called Urzymes [40].

PR is an abiotic feature of free amino acids in solution. PR is a physico-chemical property of each amino acid, defined by their migration in paper chromatographic experiments in aqueous solutions of nucleobases [41]. PR is directly related to the organization of the codon table and its amino acids [42]. In addition, PR is related to the partition of amino acid in a polar–non-polar interface [43]. The SGC is also robust to errors of single base mutations and this is reflected when PR is used as a metric of amino acid similarity [28,44,45]. Moreover, the phenotypic graphs of amino acids exhibit disjoint clusters of amino acids when their PR values are used [15]. The genetic code became optimized with respect to PR. By observing the microscopic environments of the amino acids in binary solution, it is apparent that the PR is related to how an amino acid partitions across a polar–non-polar interface. Several theoretical studies have found a high degree of error tolerance in the genetic code when PR is used as a measure of amino acid similarity [28,45–47]. Polar–non-polar interfaces may have played a role in the establishment or development of the early genetic code. It is highly improbable that the genetic code became optimized with respect to PR purely by chance.

As far as translation is concerned, it does not make sense to consider one code without the other. The present-day operational code is intricately carved in the structure of tRNA acceptors and cognate

aaRSs, whereas the anticodon code is reduced to codon–anticodon interactions. The catalytic proteins required to accelerate this binding are divided between two very ancient enzyme superfamilies, the class I and class II aaRSs, each activating 10 of the 20 canonical amino acids [8]. The present correspondence of the two codes is provided by 20 specific aaRSs divided into two strikingly dissimilar classes of 10 members each. There are only 20 aaRSs, one for each amino acid (and, respectively, for isoacceptor tRNAs); hence, the operational code is non-degenerate [12]. Such a non-degeneracy, inherent only to the acceptor code, may indicate the historically subsidiary role of anticodons in aminoacylation. Otherwise, more than 20 aaRSs could exist, one for each anticodon rather than one for each amino acid. The two aaRSs recognize the acceptor helix from opposite sides: class I aaRS approaches the helix from the side of its minor groove and attaches the amino acid to the 2'OH group of the terminal adenosine ribose, while class II aaRS approaches from the side of major groove and attaches the amino acid to the 3'OH group [8]. The aaRSs are divided into two classes distinguished by their structures [8]. The term 'class' is used to distinguish both the enzymes and the amino acids that they activate [8]. Polarity and size are used to distinguish between the two classes of amino acids [37,39]. Class II amino acids occur significantly more frequently at the surfaces of proteins, whereas class I amino acids occur more frequently in their cores [39]. Notably, the two synthetases classes seem to have descended from ancestors coded by opposite strands of the same gene [48]. There is no need for the aaRS to recognize the anticodon in order to properly aminoacylate the tRNA. This means that the two codes coevolved right at the origin of translation. This encoding system seems now lost in the dimness of the past. Rodin & Ohno [49] found that the two families of aaRSs exhibit significant sequence similarity, but only when their coding sequences are compared in the opposite direction. This finding prompted Rodin & Ohno [49] to suggest that the two synthetases families originated as two-protein coding genes located on the complementary strands of the same primordial double-stranded RNA. Assuming that the partition into two mechanisms of tRNA-aminoacylation is a relic that dates back to the primordial genetic code in the RNA world, Delarue [3] proposed a simple model based upon successive binary choices for the assignment of codons to amino acids. Both Delarue's [3] and Rodin & Rodin's [7] models reorganize the codon table to reflect these contrasting molecular recognition modes by the two aaRS classes. These authors propose that this dual complementarity is frozen from an earlier stage in the code's development, at which triplet reading frames had been established, but only the middle bases of the anticodons had been fixed, perhaps coinciding with the second step of Delarue's differentiation genealogy [7]. They concluded that new codons were recruited in pairs, because translation of both sense and antisense strands would require that meaning be attached to both codons and their anticodons. We chose these models in order to prove the power of algebraic methods to understand each model and because our approach facilitates the comparison of the predictions among different models. In particular, the RO-model has a sound experimental background [37–40,48].

Herein the RO [7,49,50] and Delarue (D) [3] models for the origin of the genetic code are analysed in terms of its symmetrical properties. The RO- and D-models are asymmetrical. In this work, we assume a primeval RNY code [9], and make the same assumption of the RO-model, i.e. that the SGC can be divided according to the two classes of aaRSs I and II. We formulate isometries with which we arrive precisely to our symmetrical algebraic model [15,21,35].

The article is organized as follows. We start with some basic definitions of group theory and we provide the definition of the group action over the set of nucleotides. Then, we analyse the Rodin–Rodin model [50] of dividing the table of the genetic code according to the two classes of aaRSs. This table is symmetric but it is biologically incorrect. Then, we formulate simple isometric transformations that allow us to transform the RO-model which is asymmetric but biologically correct, into the SGC model based on the primeval RNY code and vice versa. We define an automorphism that converts the class aaRS I into the class aaRS II. We also model the asymmetric D-model into a symmetrical one by means of quotient groups. As a direct application of the 6D model of the SGC, we used the four scales of PR of each amino acid [41] and it neatly divides the SGC into four symmetrical groups. Finally, we discuss the results in terms of our model, which is compatible with the RO- and D-models and the primeval RNY code [9]. In other words, we have a unique 6D model, which is consistent with the RNY primeval genetic code and with the distribution of the two classes of aaRS.

2. Mathematical background

Group theory is a branch of abstract algebra that deals with the notion of symmetry of a geometrical object, making the set of symmetries of an object a group structure.

Table 1. The multiplication table of the Four-Klein group (K_4, \circ) .

\circ	e	a	b	ab
e	e	a	b	ab
a	a	e	ab	b
b	b	ab	e	a
ab	ab	b	a	e

2.1. Definition of a group

A group is a set G with a binary operation \circ that combines any two elements of G and returns an element in G . This ordered pair is denoted as (G, \circ) which satisfies the following properties:

1. Closure: For all a, b in G , the resulting element is also in G .
2. Associativity: For all a, b, c in G , the next equality holds: $(a \circ b) \circ c = a \circ (b \circ c)$.
3. Identity element: There exists an element e in G such that $a \circ e = e \circ a = a$ for all a in G .
4. Inverse element: For all a in G , there exists an element a' such that $a \circ a' = a' \circ a = e$, where e is the identity element.

2.2. Definition of a group action

If G is a group and X is a set then a group action is a function $f: G \times X \rightarrow X, (a, x) \rightarrow a * x$ that satisfies the following axioms:

1. Compatibility: For all a, b in G and all x in X the equality $(a \circ b) * x = a * (b * x)$ holds.
2. Identity: For all x in $X, e * x = x$, where e is the identity element of G .

Then, it is said that G acts on X and X is a G -set.

A group action is the description of symmetries of an object using an external group. The essential elements of the object are described in a set and the operating group is known as the group of symmetries and its members correspond to some of the one-to-one transformations of the set. When considering a point $x \in X$ and the group G operating over X , the set $Gx = \{g * x | g \in G\}$ is called the orbit of the point X under the action of G . The set of orbits from a set X under the action of a group G is a partition of the set X , and it is known as the quotient set of the action, denoted by X/G .

2.3. Four-Klein group

Herein, we develop a novel and logically equivalent approach, where fewer algebraic properties are required, to that followed in our previous works [16,21,24] in which a group structure in the set $N = \{C, U, G, A\}$ of the four nucleotides was defined. Herein, the ordering of the nucleotides and their arbitrary binary assignments are no longer necessary. A group is naturally constructed with the two types of mutations, transversions and transitions, represented by a and b , respectively. These two types of transformations are used like generators of the group with the property that the composition (denoted by \circ) of a mutation with itself is equal to the identical mutation. The new approach starts with the symmetry group that corresponds to an abstract rectangle, which in group theory is known as the Four-Klein group, here symbolized as (K_4, \circ) , where $K_4 = \{e, a, b, ab = ba\}$ is the set, and \circ is the group operation (table 1). The Four-Klein group is identified as an abelian group in the direct product $\mathbb{Z}_2 \times \mathbb{Z}_2$, where $\mathbb{Z}_2 = \{0, 1\}$ represents the cyclic group of two elements. The set $\mathbb{Z}_2 \times \mathbb{Z}_2$ is regarded as the set of the four duplets of zeros and ones.

2.4. Group action in the set of nucleotides

Herein, the set of nucleotides N and its mutations will be considered. The Four-Klein group that will act over the set N , making it mutate, just as a rectangle is transformed in itself through its symmetries. This is represented as the Cayley graph of the group with the nucleotides as vertices. As an example consider the following: $a * (A) = U, a * (G) = C, b * (A) = G, b * (U) = C$, while $(a \circ b) * (A) = (b \circ a) * (A) = C$, and $(a \circ b) * (U) = (b \circ a) * (U) = G$. For the sake of simplicity, the symbols $\circ, *$ and the parentheses will be here and further omitted where no misinterpretation can be made, so that $(a \circ b) * (A) = abA$.

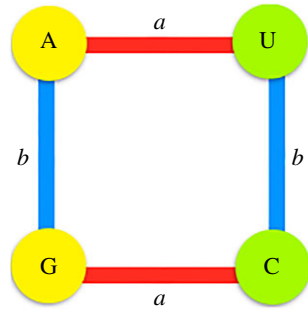


Figure 1. Representation of the action of the generators of the group over the set of nucleotides, where a represents transversions and b transitions. Purines are coloured in yellow and pyrimidines in green.

Now we extend our nucleotide level group action to the set of 64 triplets, $N \times N \times N = N^3$ as follows: $f: K_4^3 \times N^3 \rightarrow N^3, ((a_1, a_2, a_3), (x_1 x_2 x_3)) \rightarrow (a_1 x_1, a_2 x_2, a_3 x_3)$, where we have used the vector notation and f is well defined because the mapping is component-wise.

A common classification of the nucleotides can be done through their chemical properties [24]. Herein, we consider purines and pyrimidines represented as R and Y, respectively, where $R = \{A, G\}$ and $Y = \{C, U\}$. Next, we will deal with codons, which in set notation are the sets: RNY, YNR, YNY and RNR.

2.5. Defining a metric or distance in N^3

For a given choice of generators, one has to define a metric, i.e. the natural distance on the Cayley graph. Here, we have the group K_4 and its two generators a and b . The metric is defined in the following manner for x_1, x_2 in N , for single nucleotides:

1. $d(x_1, x_2) = 0$. If and only if $x_1 = ex_2$.
2. $d(x_1, x_2) = 1$. If and only if $x_1 = ax_2$ or $x_1 = bx_2$.
3. $d(x_1, x_2) = 2$. If and only if $x_1 = abx_2 = bax_2$.

This is a discrete metric that is similar to the one known as Hamming distance, but here the distance is given by the minimum number of generators of the group that are used to take one nucleotide and mutate it into another one. An extension in the definition of distance is natural for triplets so that it will be the sum of the distances of the nucleotides that conform the triplet. Formally, for two triplets $x_1 y_1 z_1$ and $x_2 y_2 z_2$, the distance is: $d(x_1 y_1 z_1, x_2 y_2 z_2) = d(x_1, x_2) + d(y_1, y_2) + d(z_1, z_2)$.

The genetic code is then represented as a 6D hypercube. This geometric figure can also be interpreted as a graph $G = (V, E)$ of vertices, representing the codons, and edges, joining the codons at distance one, making it possible to analyse its symmetries through the group of automorphisms of the graph. This group consists of all the bijective functions of the graph $G, f: (V, E) \rightarrow (V, E)$ that preserve its adjacencies. With the metric defined above, these automorphisms comprise all the isometric transformations of the cube. It is worth mentioning that there are, in essence, only three different Cayley graphs that determine the action of the group over the nucleotides. The pairs of opposite edges of the graph chosen here (figure 1) represent the generators of the group (transversions and transitions), which is in agreement with a common evolutionary interpretation [51]. In our previous approach [16,21,24], the distance of a codon and its anticodon in the 6D hypercube is at the maximum distance of 6. It is worth remarking that, if the Cayley graph associated with our previous works is used, the interchange of the action a for ab , and ab for a , applied as described above, will result in the same conclusions. Hence, the two approaches do not contradict each other, neither in biological aspects nor in mathematical ones, owing to the fact that with the present approach the ordering of nucleotides and arbitrary binary assignments are not required. In fact, the four nucleotides A,C,G,U can be situated at the vertices of a given rectangle in $4! = 24$ ways. Interestingly, the assumption that a and b represent transversion and transition, respectively, being a the transversion that converts each nucleotide into its complementary, reduces all the possible graphs to only three.

3. The Rodin–Rodin model

In the original proposal made by Rodin & Ohno [49], the table of the genetic code is arranged in such a manner, that complementary codons appear vis-à-vis each other. Each of the 20 different aaRSs

Table 2. Symmetric table of the SGC that is biologically incorrect.

	U		A		G		C				
U	Phe	U	A	Arg	A	U	Cys	U	A	Thr	A
U	Phe	C	G	Glu	A	U	Cys	C	G	Ala	A
U	Leu	A	U	Stop	A	U	Stop	A	U	Ser	A
U	Leu	G	C	Gln	A	U	Trp	G	C	Pro	A
C	Leu	U	A	Arg	G	C	Arg	U	A	Thr	G
C	Leu	C	G	Glu	G	C	Arg	C	G	Ala	G
C	Leu	A	U	Stop	G	C	Arg	A	U	Ser	G
C	Leu	G	C	Gln	G	C	Arg	G	C	Pro	G
A	Ile	U	A	Asn	U	A	Ser	U	A	Thr	U
A	Ile	C	G	Asp	U	A	Ser	C	G	Ala	U
A	Ile	A	U	Tyr	U	A	Lys	A	U	Ser	U
A	Met	G	C	His	U	A	Lys	G	C	Pro	U
G	Val	U	A	Asn	C	G	Gly	U	A	Thr	C
G	Val	C	G	Asp	C	G	Gly	C	G	Ala	C
G	Val	A	U	Tyr	C	G	Gly	A	U	Ser	C
G	Val	G	C	His	C	G	Gly	G	C	Pro	C

Table 3. Biologically correct table of the SGC that is not symmetric. Phe and Tyr are ambiguous and they are marked with an asterisk.

	U		A		G		C				
U	Phe*	U	A	Lys	A	U	Cys	U	A	Thr	A
U	Phe*	C	G	Glu	A	U	Cys	C	G	Ala	A
U	Leu	A	U	Stop	A	U	Stop	A	U	Ser	A
U	Leu	G	C	Gln	A	U	Trp	G	C	Pro	A
C	Leu	U	A	Lys	G	C	Arg	U	A	Thr	G
C	Leu	C	G	Glu	G	C	Arg	C	G	Ala	G
C	Leu	A	U	Stop	G	C	Arg	A	U	Ser	G
C	Leu	G	C	Gln	G	C	Arg	G	C	Pro	G
A	Ile	U	A	Asn	U	A	Ser	U	A	Thr	U
A	Ile	C	G	Asp	U	A	Ser	C	G	Ala	U
A	Ile	A	U	Tyr*	U	A	Arg	A	U	Ser	U
A	Met	G	C	His	U	A	Arg	G	C	Pro	U
G	Val	U	A	Asn	C	G	Gly	U	A	Thr	C
G	Val	C	G	Asp	C	G	Gly	C	G	Ala	C
G	Val	A	U	Tyr*	C	G	Gly	A	U	Ser	C
G	Val	G	C	His	C	G	Gly	G	C	Pro	C

recognizes the cognate amino acid, and then attaches it to isoacceptor tRNAs with the corresponding anticodons. The operational code provides virtually errorless aminoacylation of tRNAs [6,12,13]. The 20 aaRSs are divided into two 10-member non-overlapping classes, I and II, that have virtually nothing in common with each other as far as the primary sequence, secondary elements and 3D structures are concerned [8].

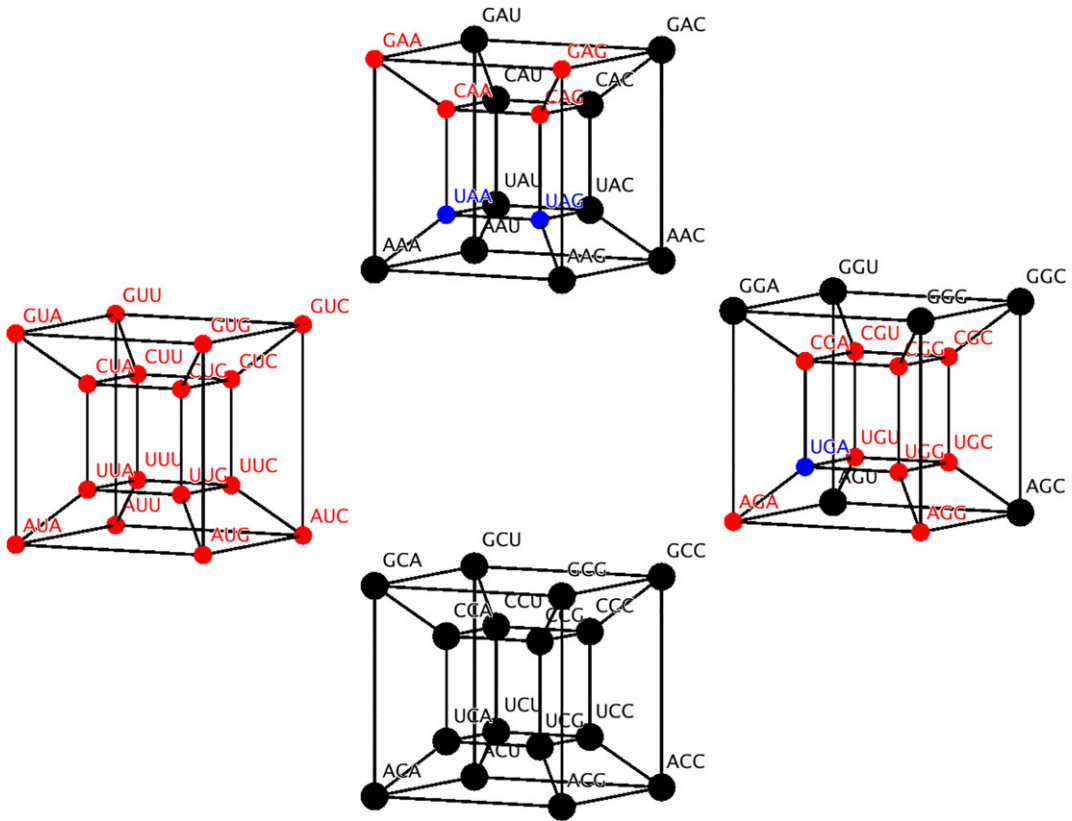


Figure 2. The six-dimensional cube of the genetic code coloured according to the aaRS class, class I is red and class II is black and bold. Stop codons (UUA, UAG and UGA) are in blue although the known cases of their ‘capture’ by amino acids are mostly from class I [52]. The edges joining the four-dimensional cube are not shown for better appreciation.

Table 4. The automorphisms used in each subcode of the SGC to interchange the aaRS classes.

T_1	T_2
RNY \leftrightarrow YNR (a,b,a)	RNY \leftrightarrow YNR (a,b,ab)
RNR \leftrightarrow RNR (e,b,e)	RNR \leftrightarrow RNR (e,b,b)
YNY \leftrightarrow YNY (e,b,e)	YNY \leftrightarrow YNY (e,b,b)

In their table, amino acids of class aaRS I are coloured in red, while those of class aaRS II are coloured in black (table 2). The amino acids from the first column of the code table tend to belong to class I (Phe being the only exception), whereas the amino acids from the second column all belong to class II.

3.1. A remarkable observation: a flaw in table 2

In table 2, there is a flaw, which conspires against the symmetries. Lysine and arginine are incorrectly placed. In arginine, two (AGA and AGG) out of its six coding triplets are incorrectly assigned to lysine, whereas the two triplets of lysine, AAA and AAG are assigned to arginine. Rodin & Rodin [50] and Rodin & Ohno [52,53] corrected table 2 [7,49], which is biologically correct but it is not symmetric (see table 3).

3.2. An automorphism that converts the class aaRS I into the class aaRS II and vice versa

The RO corrected table of codons associated to each class of aaRS lost symmetry, but in the 6D model this symmetry is recovered. Symmetries are represented with automorphisms of the cube that interchange the

Table 5. Automorphisms to convert the Rodin–Ohno model partitions of the genetic code into the RNR, RNY, YNR, YNY partitions.

_____ F _____	_____ F _____
RAR \leftrightarrow RAR <small>(e,e,e)</small>	RGR \leftrightarrow YAR <small>(a,b,e)</small>
YGR \leftrightarrow YGR <small>(e,e,e)</small>	YAR \leftrightarrow RGR <small>(a,b,e)</small>
RUY \leftrightarrow RUY <small>(e,e,e)</small>	RCY \leftrightarrow YUY <small>(a,b,e)</small>
YCY \leftrightarrow YCY <small>(e,e,e)</small>	YUY \leftrightarrow RCY <small>(a,b,e)</small>
RUR \leftrightarrow RAY <small>(e,a,a)</small>	RCR \leftrightarrow YAY <small>(a,ab,a)</small>
YCR \leftrightarrow YGY <small>(e,a,a)</small>	YUR \leftrightarrow RGY <small>(a,ab,a)</small>
RAY \leftrightarrow RUR <small>(e,a,a)</small>	RGY \leftrightarrow YUR <small>(a,ab,a)</small>
YGY \leftrightarrow YCR <small>(e,a,a)</small>	YAY \leftrightarrow RCR <small>(a,ab,a)</small>

codons of class I with class II and vice versa. In fact, there are two such functions, T_1 and T_2 defined piece-wise (table 4). These automorphisms form a subgroup that under composition yields a class invariant transformation ($T_1 \circ T_2$) = $T_3 = (e, e, b)$, which is a transition in the wobble position. In figure 2, the codons in the 6D model are coloured according to the aaRS class as in table 2 and table 3 but black is replaced by blue. Each isolated cube is actually a four-dimensional (4D) cube and the union of all of them with their respective extra edges forms the complete 6D cube. The edges joining each 4D cube are omitted for a better appreciation of the complete figure.

3.3. From the RO-model to the standard genetic code

According to the RO-model [49], the table of the genetic code can be divided into the sub-codes NAN, NGN, NUN, NCN. There exists an automorphism F of the cube defined also piece-wise, which transforms that division into the sub-codes RNR, YNR, RNY, YNY, respectively (table 5), which is precisely our algebraic model [16,21,35]. As an example, consider the codon AGC in the RO-model. AGC is an element of the RGY subcode, so the action required to transform it to our 6D model is (a,ab,a) as described in table 5. From the definition of the group action, this codon will be transformed to the triplet UUG. Note also that, owing to the order of the elements of the group, the same action over UUG on the 6D model will send it back to AGC in the RO-model.

4. The polar requirement in the six-dimensional SGC

PR was scaled into four categories [41]. We assign a particular colour (red, yellow, blue and green) to each scale. When such categories are set on the 6D genetic code, new symmetries emerge (figure 3). Now the SGC in six dimensions can be symmetrically divided into four colours according to the PR. Each category, or colour, comprises 16 codons that are arranged in 4D hypercubes, whose symmetry is given by the wreath product S_2WrS_4 , where S_n is a permutation group of n elements [54]. Such group can be represented by the group of orthogonal matrices of 4×4 whose entries are all integers [54]. To interchange whole categories, it is sufficient to use the symmetries of a square Dih_4 (figure 3). Hence, the 6D representation of the SGC can reflect this property using its automorphisms as a biological classifier.

4.1. Delarue's model

Delarue [3] argues that the partition of codons according to the aaRS class distinction facilitated a hierarchical process by which additions to the code reduced codon ambiguity to produce the extant table with just five binary choices. The code started with undifferentiated and nonsense triplets, NNN. Codons were given meaning beginning with the second base and ending with the third. The NYN triplets could interact with a synthetase, whereas the NRN could not and remained stop codons. At each step, the ambiguous codon family differentiated to give descendants with opposite groove recognition, while descent of the stop codon family generated a new ambiguous family and retained a stop codon,

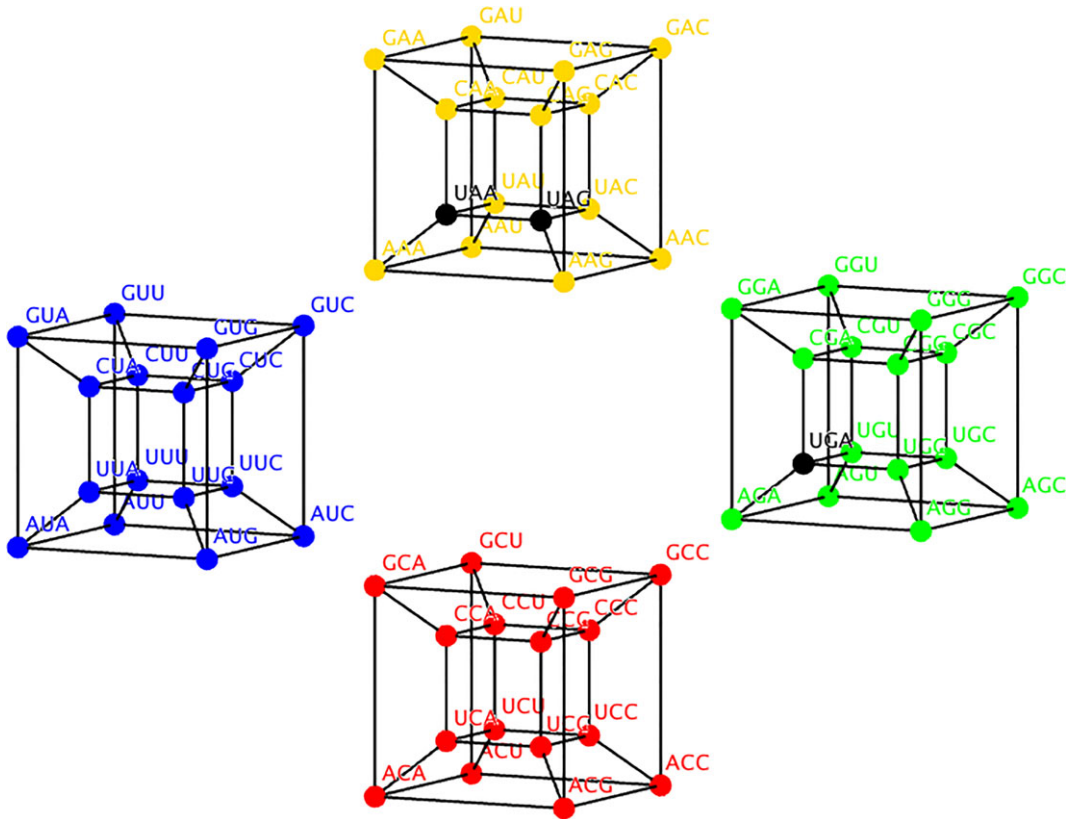


Figure 3. Six-dimensional hypercube of the SGC coloured by amino acid polar requirement values [41]. The four-dimensional hypercubes are yellow (upper); blue (left); red (lower); green (right); Stop codon are in black (UUA, UAG and UGA).

which was always present in the code. These asymmetric division rules provide a unique differentiation order, rendering the exhaustive exploration of the initial assignment of codons plausible, and suggesting that the appearance of the code conferred meaning successively from redundancy by a deterministic elimination of the most frequent errors. Notably, tRNAs with complementary anticodons also have statistically significant complementarity in their acceptor-stem operational codes [3].

With the concept of group action in mind, it is possible to analyse the D-model and elaborate an algebraic model. As the order 2 subgroup T generated by b is the group of transitions of the set N , $T = \{e, b\}$ is isomorphic to the cyclic group \mathbb{Z}_2 . The quotient N/T represents precisely the partitions $\{R, Y\}$ of the set of nucleotides. Considering the quotient N/e , where e is the trivial group, we obtain the nucleotides separated in different sets: $N/e = \{\{A\}, \{G\}, \{U\}, \{C\}\}$. Finally, the quotient with the entire group is a trivial operation, with only one class, as $N/K_4 = N$. In order to analyse triplets, a component-wise operation naturally arises from these definitions resulting in: $NNN/GGG = N/G \times N/G \times N/G$, where G is any subgroup of K_4 , i.e. analysing the quotients component by component and then relating them with the cartesian products of sets. Now the Delarue's model given by six binary choices can be algebraically analysed. The quotient $NNN/K_4TK_4 = \{NRN, NYN\}$, where $T = \{e, b\}$ is the subgroup of transitions, yields the first binary choice and for the next steps doing NRN/K_4eK_4 and NYN/K_4eK_4 , respectively, and for the rest what is only needed is to use as quotients the products: TeK_4 , eeK_4 , eeT and eee in that order. We have just replaced the six binary decisions (including the wobbling assignments) in the D-model by six algebraic well-defined mathematical representations. The value of the latter is that we can follow the groups of symmetries in each step. Furthermore, we can make the model parsimonious and simpler, if we now make quaternary decisions so that the nucleotides in each position of the codon are determined at each step, by the use of only three group products, K_4eK_4 , eeK_4 and eee .

5. Discussion

In this work, we have been able to formulate algebraic expressions for two well-known models of the origin and evolution of the genetic code, to wit, the RO-model and Delarue's model. Both models are

consistent with the RNY code [9], as partitioning of aaRSs in two classes could have been encoded in a strand-symmetric RNA world [7,50]. We have shown that by assuming both a primeval RNY code and that the code can be divided into two classes of the aaRSs, we arrive at a symmetrical representation of the genetic code in a 6D algebraic model. We have also shown that PR displays a symmetrical pattern in this 6D model. PR is an empirical scale unrelated to either of the two transfer equilibria that best represent the partitioning of amino acids between pure phases, rather than between a pure phase and cellulose. PR seems to be also unrelated to other measures such as hydrophathy. Further experimental work is needed to clarify these issues.

The aaRSs are a prime example of horizontal gene transfer [55,56]. Evolutionary replacements of aaRSs accompanied the evolution of the genetic code [31]. The assignments seemed to minimize errors in a primitive translation mechanism that was highly inaccurate [57,58]. The evolutionary phylogenies of synthetases do not obey the basic division of all life into the three primary groupings Bacteria, Archaea and Eukaryotes [56]. The two aaRS classes are presumably the oldest protein superfamilies. The RO hypothesis [52] implies that they arose at nearly the same instant in geological time because, at the nucleic acid level, the information necessary for function of each class is indistinguishable from that necessary for function of the other [40]. Complementarity means that one strand implies the existence of the other. Sense/antisense coding thus projects back past the genetic coding nexus to chemistry. The sense/antisense ancestry of the aaRS appears to be solidly established [40,59]. The authors, Rodin & Ohno, observed that their model is *almost perfectly symmetric* [49,52,53]. But in front of this unusual assertion we argue that something that is almost perfectly symmetric is not symmetric at all. Interestingly, the automorphisms T_1 and T_2 show the so-called symmetry that only exists in our 6D model, and the function F converts the partitions of the RO-model {NUN, NAN, NGN, NCN} into the partition {RNR, RNY, RNY, YNR}, which corresponds to our symmetric model. As the functions presented are isometric, the RO-model may be considered as equivalent to this one and it only takes a different point of view of the same model to reach one's conclusions from the other. The D-model is a phenomenological model of progressive differentiation-like reduction of codon ambiguity [60]. Indeed, it has been suggested that the primitive ribosome worked to synthesize peptides randomly, without the need of a code [61]. This elegant model is also based on the pattern of tRNA aminoacylation by class I and II aaRSs. However, in contrast with our complementarity-based model, Delarue's asymmetric model consists of a binary decision tree, like in a longitudinal differentiation process [3]. The whole SGC is derived from binary decisions but it remains unclear why the minor or major groove side is preferred in each particular step. We propose an algebraic model that accounts for the simultaneous selection of pairs of complementary triplets following the RO-model, and a set of six algebraic well-defined algebraic operations that account for the six binary decisions of the D-model. We have shown that the D-model can be built from simple operations of action groups. The preservation of symmetries is noteworthy. With only two transformations, we can derive, from a single codon, the 32 triplets forming the RNY and YNR subsets, as well as the 32 triplets comprising the sets RNR and YNY. All the transformations required for the construction are subgroups of K_4^3 which is the general group acting on the codon space, therefore making impossible the creation of new codons without a symmetry breaking which is the action of a new subset of operators.

Until now, participation of two aaRS classes in genetic coding has been rationalized as a result of successive binary choices [3] or as a means of avoiding coding ambiguity [60]. It has been shown that this distinction appears to be related to the complementary roles of class I and II amino acids in protein folding. Members of subclass IA (Leu, Ile, Val and Met) have aliphatic side-chains and are found in hydrophobic cores. Members of subclass IIA (Ser, Thr and His) are small amino acids with water-favouring side-chains. Subclasses B (with carboxyl, amide, primary amine side-chains) and C (aromatic) in both classes contain similar amino acids. Class I amino acids tend to be buried; those in class II remain largely on the surface. Class I amino acids allowed formation of non-polar cores and class II amino acids populated the surfaces of globular proteins. The linkage between classes arising from their sense/antisense ancestry [38,62] would be expected to simplify the search for reduced amino acid alphabets that may have been used during early protein evolution, leading to the universal genetic code. The order in which predictors emerge in the stepwise regressions discussed above is similar, but not identical to, the series of decisions by which Delarue suggested that genetic coding actually became fixed [3]. Although tRNA identity elements have probably been confounded by horizontal gene transfer [32], ancestral tRNA sequence reconstruction may clarify further how identity elements and the synthetase class recognition evolved.

With our approach, we have shown that the whole SGC can be derived starting from a pair of reverse complementary codons with just six steps or just three if we follow quaternary decisions. The present

algebraic approach is general and abstract enough as it deals with the algebra from outside of the genetic code making it possible to build bridges among different models. This approach permits the direct comparison of different genetic models that otherwise would be difficult to perform. For example, the self-referential (SR) model for the formation of the SGC [14] is appealing because it considers a self-modifying genetic code that alters its own instructions while it is evolving. Consequently, the instruction path length is reduced and improves its performance and maintenance through the mechanism of natural selection. It is called SR because it is centred on the integration of self-feeding ribonucleoprotein structures where the protein and RNA activities are mutually stimulatory, after having been formed on top of the basic tRNA dimers. It assumes that during early stages of the formation of the SGC, protein synthesis was directed by tRNA dimers. The SR-model lacks experimental support but it is compatible with the appearance of the metabolic pathways [63]. The proposed dimer-directed transferase activity should be experimentally tested, either utilizing present-day tRNAs or the various kinds of mini-tRNAs that have been used as acceptors for the aaRS function or for spontaneous aminoacylation. The genetic eukaryotic anticodon comprises 46 anticodons as there are not anticodons ending with adenine ($3' \rightarrow 5'$) direction. The group actions required to describe the symmetries of this model are given by the direct product $K_4 \times K_4 \times \mathbb{Z}_3$, where the last set is the cyclic group of three elements that corresponds to the rotations of a triangle. The cyclic groups are generated with one element so the biological interpretation of this action is ambiguous, in contrast with the generators of K_4 representing transitions and transversions. Another difference is that this model can only be fully described in five dimensions. These differences in the mathematical properties of the SR-model with our 6D model show that they are non-equivalent and that there is no smooth way to mathematically complete the SGC. Essentially, the problem lies in the fact that the group K_4 cannot be obtained from \mathbb{Z}_3 . The SR-model lacks an explanation of how the dinucleotides formed the codons. Did they appear gradually? Or did codons appear simultaneously from a given set of principal dinucleotides? The chronology of appearance of codons is absent.

The partition of the table of the genetic code into the two classes of aaRSs is entirely consistent with the complementary symmetry of the RNA world in general, and the hypothesis of its initial double-strand coding in particular. It has been shown that the elimination of any amino acid encoded by the primeval RNY code would be strongly selected against and therefore at this stage the RNY code was already frozen [18]. The very existence of the ying-yang (formerly dubbed 'ying-yang-like' [7]) pattern of aminoacylation that certainly has little if anything at all to do with the present-day protein aaRSs, points to the 'anticodon first' scenario of the genetic code origin [64,65]. The anticodon is indeed essential for 17 of the 20 *Escherichia coli* isoaccepting groups [66]. The second operational code does not make sense without the anticodon code. However, the early relevance of the acceptor mini-helix in evolutionary development of the tRNA molecule cannot be understated [13,36,52,59,67,68]. Consistent with the hypothesis that the acceptor double-stranded stem is older than the anticodon loop, the GC-biased codon-anticodon-like triplet pairs located just next to the 73rd base-determinator of the acceptor stem may better reflect the very initial shaping stage of the genetic code than the single-stranded anticodon [50,53].

We have developed mathematical models for the RO-hypothesis and the D-genealogy. We highlight that these mathematical models are different despite the fact that they share the fundamental fact that the SGC can be divided by the two classes of aaRSs. We emphasize that our 6D model is completely equivalent to the mathematical model of the RO-hypothesis. The mathematical model of the SR-hypothesis underscores the differences with the other three models. The 6D symmetrical model has been enriched by the RO-model and the RO-model has acquired a sound mathematical structure. All presented models deal with the same biological aspects of the SGC, but differently. The 6D structure has been exploited not only for comparing different models but more importantly to give a step forward to unify models and reinforce (or weaken) models' hypotheses.

In conclusion, the most adequate model for the SGC can be represented in a 6D hypercube. Each dimension describes a type of mutation, transition and transversion as given by the Cayley graph, acting on each of three bases of any codon. Consequently, we obtain the six dimensions. When considering the hydropathy scale of amino acids [69], there are no symmetries that would interchange the four categories. However, if the codon UGA were associated with an amino acid that falls into the category of 'moderately hydrophobic', then the transformation (e, e, b) would be invariant to the hydropathy classes. In the same manner, when considering the polarity of amino acids [37–39], it would be needed that UGA were a non-polar amino acid, the transformation (e, e, b) would be invariant to polarity. If, in addition, the other stop codons are assigned to polar amino acids, the transformation (e, e, a) would be another invariant symmetry, as well as their composition. This means that a biological classification can also be interpreted as symmetries that would maintain the classification.

Undoubtedly, the 6D description of the genetic code as the hypercube (\mathbb{Z}_2)⁶, becomes essential for a better understanding of the evolution of the code. The SGC, as derived from the primeval genetic code, and the RO-model are one and the same. We have shown that these different models of the genetic code are mathematically equivalent. Hence, the 6D algebraic model presented here unifies different models of the genetic code.

Data accessibility. The values of polar requirement were taken from [41]; the hydropathy scale of amino acids were taken from [69]. The rest of the work is essentially theoretical.

Authors' contributions. M.V.J. conceived the whole work, performed the analysis, interpretation, coordinated the research and wrote the manuscript; G.S.Z. and E.R.M. contributed to ideas, design, calculations, and wrote drafts of the manuscript. All authors approved the final version to be published.

Competing interests. The authors declare no competing interests.

Funding. M.V.J. was financially supported by PAPIIT-IN224015, UNAM, México. G.S.Z. is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and received fellowship number: 737920 from CONACYT.

Acknowledgements. We thank Juan R. Bobadilla for the technical computer support.

References

- Darwin C. 1859 *The origin of species: by means of natural selection, or the preservation of favoured races in the struggle for life*. Cambridge, UK: Cambridge University Press.
- Nirenberg MW, Matthaei JH. 1961 The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc Natl Acad. Sci. USA* **47**, 1588–1602. (doi:10.1073/pnas.47.10.1588)
- Delarue M. 2008 An asymmetric underlying rule in the assignment of codons: possible clue to a quick early evolution of the genetic code via successive binary choices. *RNA* **13**, 161–169. (doi:10.1261/rna.257607)
- Crick FHC. 1968 The origin of the genetic code. *J. Mol. Biol.* **38**, 367–379. (doi:10.1016/0022-2836(68)90392-6)
- Crick FHC, Brenner S, Klug A, Piecznik GA. 1976 Speculation on the origin of protein synthesis. *Orig. Life* **7**, 389–397. (doi:10.1007/BF00927934)
- Ribas de Pouplana L, Schimmel P. 2001 Aminoacyl-tRNA synthetases: potential markers of genetic code development. *Trends Biochem. Sci.* **26**, 591–596. (doi:10.1016/S0968-0004(01)01932-6)
- Rodin SN, Rodin SA. 2008 On the origin of the genetic code: signatures of its primordial complementarity in tRNAs and aminoacyl-tRNA synthetases. *Heredity* **100**, 341–355. (doi:10.1038/sj.hdy.6801086)
- Eriani G, Delarue M, Poch O, Gangloff J, Moras D. 1990 Partition of aminoacyl-tRNA synthetases into two classes based on mutually exclusive sets of conserved motifs. *Nature* **347**, 203–206. (doi:10.1038/347203a0)
- Eigen M, Schuster P. 1978 The hypercycle: a principle of natural selection. *Naturwissenschaften* **65**, 341–369. (doi:10.1007/BF00439699)
- Eigen M, Winkler-Oswatitsch R. 1981 Transfer-RNA: the early adaptor. *Naturwissenschaften* **68**, 217–228. (doi:10.1007/BF01047323)
- Eigen M, Winkler-Oswatitsch R. 1981 Transfer-RNA, an early gene? *Naturwissenschaften* **68**, 282–292. (doi:10.1007/BF01047470)
- De Duve C. 1988 The second genetic code. *Nature* **333**, 117–118. (doi:10.1038/333117a0)
- Schimmel P, Giégé R, Moras D, Yokoyama S. 1993 An operational RNA code for amino acids and possible relationship to genetic code. *Proc. Natl Acad. Sci. USA* **90**, 8763–8768. (doi:10.1073/pnas.90.19.8763)
- Guimarães RC, Costa Moreira CH, Farias ST. 2008 AP self-referential model for the formation of the genetic code. *Theory Biosci.* **127**, 249–270. (doi:10.1007/s12064-008-0043-y)
- José MV, Morgado ER, Guimarães RC, Zamudio GS, Fariás ST, Bobadilla JR, Sosa D. 2014 Three-dimensional algebraic models of the tRNA code and the 12 graphs for representing the amino acids. *Life* **4**, 341–373. (doi:10.3390/life4030341)
- José MV, Morgado ER, Govezensky T. 2007 An extended RNA code and its relationship to the standard genetic code: an algebraic and geometrical approach. *Bull. Math. Biol.* **69**, 215–243. (doi:10.1007/s11538-006-9119-3)
- Novozhilov AS, Wolf YI, Koonin E. 2007 Evolution of the genetic code: partial optimization of a random code for robustness to translation error in a rugged fitness landscape. *BM Cent. Biol. Dir.* **2**, 1–24. (doi:10.1186/1745-6150-2-1)
- José MV, Zamudio GS, Palacios-Pérez M, Bobadilla JR, Fariás ST. 2015 Symmetrical and thermodynamic properties of phenotypic graphs of amino acids encoded by the primeval RNY code. *Orig. Life Evol. Biosph.* **45**, 77–83. (doi:10.1007/s11084-015-9427-4)
- Lewin B. 2000 *Genes* (vol. VII). New York, NY: Oxford University Press.
- Crick FHC. 1966 Genetic code: yesterday, today and tomorrow. *Cold Spring Harb. Symp. Quant. Biol.* **31**, 1–5. (doi:10.1101/SQB.1966.031.01.006)
- José MV, Morgado ER, Govezensky T. 2011 Genetic hotels for the standard genetic code: evolutionary analysis based upon novel three-dimensional algebraic models. *Bull. Math. Biol.* **73**, 1443–1476. (doi:10.1007/s11538-010-9571-y)
- Sánchez R, Grau R, Morgado E. 2006 A novel Lie algebra of the genetic code over the Galois field of four DNA bases. *Math. Biosci.* **202**, 156–174. (doi:10.1016/j.mbs.2006.03.017)
- Jiménez-Montaña MA, de la Mora-Basañez CR, Pöschel T. 1996 The hypercube structure of the genetic code explains conservative and non-conservative amino acid substitutions *in vivo* and *in vitro*. *Biosystems* **39**, 117–125. (doi:10.1016/0303-2647(96)01605-X)
- José MV, Morgado ER, Sánchez R, Govezensky T. 2012 The 24 possible algebraic representations of the standard genetic code in six and three dimensions. *Adv. Stud. Biol.* **4**, 119–152.
- Arquès DG, Michel CJ. 1996 A complementary circular code in the protein coding genes. *J. Theor. Biol.* **182**, 45–58. (doi:10.1006/jtbi.1996.0142)
- Michel CJ, Pirillo G, Pirillo MA. 2008 A relation between trinucleotide comma-free codes and trinucleotide circular codes. *J. Theor. Biol.* **401**, 17–26. (doi:10.1016/j.tics.2008.02.049)
- Pohlmeier R. 2008 The genetic code revisited. *J. Theor. Biol.* **253**, 623–624. (doi:10.1016/j.jtbi.2008.04.028)
- Freeland SJ, Hurst LD. 1998 The genetic code is one in a million. *J. Mol. Evol.* **47**, 238–248. (doi:10.1007/PL00006381)
- Eigen M *et al.* 1989 How old is the genetic code? Statistical geometry of tRNA provides an answer. *Science* **244**, 673–679. (doi:10.1126/science.2497522)
- Nicholas HB, McClain WH. 1995 Searching tRNA sequences for relatedness to aminoacyl-tRNA synthetase families. *J. Mol. Evol.* **40**, 482–486. (doi:10.1007/BF00166616)
- Nagel GM, Doolittle RF. 1995 Phylogenetic analysis of aminoacyl-tRNA synthetases. *J. Mol. Evol.* **40**, 487–498. (doi:10.1007/BF00166617)
- Woese CR, Olsen GJ, Ibba M, Söll D. 2000 Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol. Mol. Biol. Rev.* **64**, 202–236. (doi:10.1128/MMBR.64.1.202-236.2000)
- Hornos JEM, Hornos YMM. 1993 Algebraic model for the evolution of the genetic code. *Phys. Rev. Lett.* **71**, 4401–4404. (doi:10.1103/PhysRevLett.71.4401)
- Sánchez R, Morgado ER, Grau R. 2005 A genetic code boolean structure, I: the meaning of Boolean deductions. *Bull. Math. Biol.* **67**, 1–14. (doi:10.1016/j.bulm.2004.05.005)
- José MV, Govezensky T, García JA, Bobadilla JR. 2009 On the evolution of the standard genetic code: vestiges of scale invariance from the RNA World in current prokaryote genomes. *PLoS ONE* **4**, e4340. (doi:10.1371/journal.pone.0004340)
- Schimmel P. 1995 An operational RNA code for amino acids and variations in critical nucleotide

- sequences in evolution. *J. Mol. Evol.* **40**, 531–536. (doi:10.1007/BF00166621)
37. Wolfenden R, Lewis CA, Yuan Y, Carter Jr CW. 2015 Temperature dependence of amino acid hydrophobicities. *Proc. Natl Acad. Sci. USA* **112**, 7484–7488. (doi:10.1073/pnas.1507565112)
 38. Carter Jr CW, Wolfenden R. 2015 tRNA acceptor stem and anticodon bases form independent codes related to protein folding. *Proc. Natl Acad. Sci. USA* **112**, 7489–7494. (doi:10.1073/pnas.1507569112)
 39. Carter Jr CW, Wolfenden R. 2016 tRNA acceptor-stem and anticodon bases embed separate features of amino acid chemistry. *RNA Biol.* **13**, 145–151. (doi:10.1080/15476286.2015.1112488)
 40. Carter Jr CW *et al.* 2014 The Rodin–Ohno hypothesis that two enzyme superfamilies descended from one ancestral gene: an unlikely scenario for the origins of translation that will not be dismissed. *Biol. Direct* **9**, 11. (doi:10.1186/1745-6150-9-11)
 41. Woese CR, Dugre DH, Saxinger WC, Dugre SA. 1966 The molecular basis for the genetic code. *Proc. Natl Acad. Sci. USA* **55**, 966–974. (doi:10.1073/pnas.55.4.966)
 42. Alff-Steinberger C. 1969 The genetic code and error transmission. *Proc. Natl Acad. Sci. USA* **64**, 584–591. (doi:10.1073/pnas.64.2.584)
 43. Mathew DC, Luthey-Schulten Z. 2008 On the physical basis of the amino acid polar requirement. *J. Mol. Evol.* **66**, 519–528. (doi:10.1007/s00239-008-9073-9)
 44. Freeland SJ, Knight RD, Landweber LF, Hurst LD. 2000 Early fixation of an optimal genetic code. *Mol. Biol. Evol.* **17**, 511–518. (doi:10.1093/oxfordjournals.molbev.a026331)
 45. Haig D, Hurst LD. 1991 A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.* **33**, 412–417. (doi:10.1007/BF02103132)
 46. Caporaso JG, Yarus M, Knight R. 2005 Error minimization and coding triplet/binding site associations are independent features of the canonical genetic code. *J. Mol. Evol.* **61**, 597–607. (doi:10.1007/s00239-004-0314-2)
 47. Di Giulio M. 1989 The extension reached by the minimization of the polarity distances during the evolution of the genetic code. *J. Mol. Evol.* **29**, 288–293. (doi:10.1007/BF02103616)
 48. Martínez-Rodríguez L *et al.* 2015 Functional class I and II amino acid-activating enzymes can be coded by opposite strands of the same gene. *J. Biol. Chem.* **290**, 19 710–19 725. (doi:10.1074/jbc.M115.642876)
 49. Rodin SN, Ohno S. 1995 Two types of aminoacyl-tRNA synthetases originally encoded by complementary strands of the same nucleic acid. *Orig. Life Evol. Biosph.* **25**, 565–589. (doi:10.1007/BF01582025)
 50. Rodin SN, Rodin SA. 2006 Partitioning of aminoacyl-tRNA synthetases in two classes could have been encoded in a strand-symmetric RNA world. *DNA Cell Biol.* **25**, 617–626. (doi:10.1089/dna.2006.25.617)
 51. Kimura M. 1981 Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl Acad. Sci. USA* **78**, 454–458. (doi:10.1073/pnas.78.1.454)
 52. Rodin S, Rodin A, Ohno S. 1996 The presence of codon–anticodon pairs in the acceptor stem of tRNAs. *Proc. Natl Acad. Sci. USA* **93**, 4537–4542. (doi:10.1073/pnas.93.10.4537)
 53. Rodin SN, Ohno S. 1997 Four primordial modes of tRNA synthetase recognition, determined by the (G,C) operational code. *Proc. Natl Acad. Sci. USA* **94**, 5183–5188. (doi:10.1073/pnas.94.10.5183)
 54. Young A. 1930 On quantitative substitutional analysis 5. *Proc. Lond. Math. Soc. Second Ser.* **31**, 273–288. (doi:10.1112/plms/s2-31.1.273)
 55. Wolf YI, Aravind L, Grishin NV, Koonin EV. 1999 Evolution of aminoacyl-tRNA synthetases: analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfers. *Genet. Res.* **9**, 689–710.
 56. Woese CR, Kandler O, Wheelis ML. 1990 Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl Acad. Sci. USA* **87**, 4576–4579. (doi:10.1073/pnas.87.12.4576)
 57. Woese CR. 1965 On the evolution of the genetic code. *Proc. Natl Acad. Sci. USA* **54**, 1546–1552. (doi:10.1073/pnas.54.6.1546)
 58. Woese CR. 1973 Evolution of the genetic code. *Naturwissenschaften* **60**, 447–459. (doi:10.1007/BF00592854)
 59. Rodin AS, Rodin SN, Carter Jr CW. 2009 On primordial sense–antisense coding. *J. Mol. Evol.* **69**, 555–567. (doi:10.1007/s00239-009-9288-4)
 60. Carter Jr CW. 2008 Thawing the ‘Frozen Accident’. *Heredity* **100**, 339–340. (doi:10.1038/hdy.2008.7)
 61. Belousoff MJ, Davidovich C, Bashan A, Yonath A. 2010 On the development towards the modern world: a plausible role of uncoded peptides in the RNA world. In *Origins of life and evolution of biospheres* (eds K Ruiz-Mirazo, PL Luisi), pp. 415–419. Berlin, Germany: Springer.
 62. Chandrasekaran SN, Yardimici GG, Erdogan O, Roach J, Carter Jr CW. 2013 Statistical evaluation of the Rodin–Ohno hypothesis: sense/antisense coding of ancestral class I and II aminoacyl-tRNA synthetases. *Mol. Biol. Evol.* **30**, 1588–1604. (doi:10.1093/molbev/mst070)
 63. Guimarães RC. 2011 Metabolic basis for the self-referential genetic code. *Orig. Life Evol. Biosph.* **41**, 357–371. (doi:10.1007/s11084-010-9226-x)
 64. Rodin AS, Szathmáry E, Rodin SN. 2011 On origin of genetic code and tRNA before translation. *Biol. Direct* **6**, 14. (doi:10.1186/1745-6150-6-14)
 65. Szathmáry E. 1991 Codon swapping as a possible evolutionary mechanism. *J. Mol. Evol.* **32**, 178–182. (doi:10.1007/BF02515390)
 66. Saks MS, Sampson JR, Abelson JN. 1994 The transfer RNA identity problem: a search for rules. *Science* **263**, 191–197. (doi:10.1126/science.7506844)
 67. Fox GE, Naik AK. 2004 The evolutionary history of the translation machinery. In *The genetic code and the origin of life* (ed. LR de Pouplana), pp. 92–105. New York, NY: Landes Bioscience.
 68. Maizels N, Weiner AM. 1994 Phylogeny from function: evidence from the molecular fossil record that tRNA originated in replication, not translation. *Proc. Natl Acad. Sci. USA* **91**, 6729–6734. (doi:10.1073/pnas.91.15.6729)
 69. Farias ST, Costa Moreira CH, Guimarães RC. 2007 Structure of the genetic code suggested by the hydrophathy correlation between anticodons and amino acid residues. *Orig. Life Evol. Biosph.* **37**, 83–103. (doi:10.1007/s11084-006-9008-7)

Apéndice 2

Article

On the Uniqueness of the Standard Genetic Code

Gabriel S. Zamudio and Marco V. José *

Theoretical Biology Group, Instituto de Investigaciones Biomédicas,
Universidad Nacional Autónoma de México, México D.F. 04510, Mexico; gazaso92@gmail.com

* Correspondence: marcojose@biomedicas.unam.mx; Tel.: +52-5562-3894

Academic Editor: Koji Tamura

Received: 15 December 2016; Accepted: 8 February 2017; Published: 13 February 2017

Abstract: In this work, we determine the biological and mathematical properties that are sufficient and necessary to uniquely determine both the primeval RNY (purine-any base-pyrimidine) code and the standard genetic code (SGC). These properties are: the evolution of the SGC from the RNY code; the degeneracy of both codes, and the non-degeneracy of the assignments of aminoacyl-tRNA synthetases (aaRSs) to amino acids; the wobbling property; the consideration that glycine was the first amino acid; the topological and symmetrical properties of both codes.

Keywords: RNY code; Standard genetic code; evolution of the genetic code; frozen code; degeneracy; aminoacyl-tRNA synthetases; symmetry

1. Introduction

A fundamental feature of all life forms existing on Earth is that, with several minor exceptions, they share the same standard genetic code (SGC). This universality led Francis Crick to propose the frozen accident hypothesis [1], i.e., the SGC does not change. According to Crick [1], the SGC code remained universal because any change would be lethal, or would have been very strongly selected against and extinguished.

The astonishing diversity of living beings in the history of the biosphere has not been halted by a frozen SGC. The inherent structure of the frozen SGC, in concert with environmental influences, has unleashed life from determinism.

It is widely accepted that there was an age in the origin of life in which RNA played the role of both genetic material and the main agent of catalytic activity [1–3]. This period is known as the RNA World [4,5].

The reign of the RNA World on Earth probably began no more than about 4.2 billion years ago, and ended no less than about 3.6 billion years ago [6]. Eigen and coworkers (1968) [7] revealed kinship relations by alignments of tRNA sequences and they concluded that the genetic code is not older than but almost as old as our planet. There is an enormous leap from the RNA World to the complexity of DNA replication, protein manufacture and biochemical pathways. Code stability since its formation on the early Earth has contributed to preserving evidence of the transition from an RNA World to a protein-dependent world.

The transfer RNA (tRNA) is perhaps the most important molecule in the origin and evolution of the genetic code. Just two years after the discovery of the double-helix structure of DNA, Crick [8,9] proposed the existence of small adaptor RNA molecules that would act as decoders carrying their own amino acids and interacting with the messenger RNA (mRNA) template in a position for polymerization to take place.

The SGC is written in an alphabet of four letters (C, A, U, G), grouped into words three letters long, called triplets or codons. Crick represented the genetic code in a two-dimensional table arranged in such a way that it is possible to readily find any amino acid from the three letters, written in the 5'

to 3' direction of the codon [1]. Each of the 64 codons specifies one of the 20 amino acids or else serves as a punctuation mark signaling the end of a message.

Crick proposed the wobble hypothesis [10,11], which accounts for the degeneracy of the SGC: the third position in each codon is said to wobble because it is much less specific than the first and second positions.

Given 64 codons and 20 amino acids plus a punctuation mark, there are $21^{64} \approx 4 \times 10^{84}$ possible genetic codes. This staggering number is beyond any imaginable astronomical number, the total count of electrons in the universe being well below this number. Note, however, that this calculation tacitly ignores the evolution of the SGC. If we assume two sets of 32 complementary triplets where each set codes for 10 amino acids, we would have $10^{32} \times 10^{32} = 10^{64}$ possible codes. Then we have a reduction of the order of 4×10^{20} . Albeit this is a significant reduction, it is still a very large number. Many more biological constraints are necessary. The result that only one in every million random alternative codes is more efficient than the SGC [12] implies that there could be $\sim 4 \times 10^{78}$ genetic codes as efficient as the SGC. This calculation does not offer deeper insights concerning the origin and structure of the SGC, particularly the frozen accident.

Crick [1] argued that the SGC need not be special at all; it could be nothing more than a “frozen accident”. This concept is not far away from the idea that there was an age of miracles. However, as we show in this article, there are indeed several features that are special about the SGC: first, it can be partitioned into two classes of aminoacyl-tRNA synthetases (aaRSs) [13]; secondly, the SGC can be broken down into a product of simpler groups reflecting the pattern of degeneracy observed [14,15]; third, it has symmetrical properties, and evolution did not erase its own evolutionary footsteps [16].

Several models on the origin of the genetic code from prebiotic constituents have been proposed [17–21]. Among the 20 canonical amino acids of the biological coding system, the amino acid glycine is one of the most abundant in prebiotic experiments that simulate the conditions of the primitive planet, either by electrical discharges or simulations of volcanic activity [22–24], and this amino acid is also abundant in the analysis of meteorites [25]. Bernhardt and Patrick (2014), and Tamura (2015) [26,27] also suggested that glycine was the first amino acid incorporated into the genetic code according to an internal analysis of its corresponding tRNA and its crucial importance in the structure and function of proteins. Part of this abundance can be ascribed to its structural simplicity when compared with the structure of the remaining 19 canonical amino acids. Several models for the origin of the coding system mirror glycine as one of the initial amino acids in this system [26–29].

The SGC was theoretically derived from a primeval RNY (R means purine, Y pyrimidine, and N any of them) genetic code under a model of sequential symmetry breakings [14,15], and vestiges of this primeval RNY genetic code were found in current genomes of both Eubacteria and Archaea [16]. All distance series of codons showed critical-scale invariance not only in RNY sequences (all ORFs (Open reading frames) concatenated after discarding the non-RNY triplets), but also in all codons of two intermediate steps of the genetic code and in all kind of codons in the current genomes [16]. Such scale invariance has been preserved for at least 3.5 billion years, beginning with an RNY genetic code to the SGC throughout two evolutionary pathways. These two likely evolutionary paths of the genetic code were also analyzed algebraically and can be clearly visualized in three, four and six dimensions [15,30,31].

The RNY subcode is widely considered as the primeval genetic code [32]. It comprises 16 triplets and eight amino acids, where each amino acid is encoded by two codons. The abiotic support of the RNY primeval code is in agreement with observations on abundant amino acids in Miller's sets [33] and in the chronology of the appearance of amino acids according to Trifonov's review [34]. It has been shown that once the primeval genetic code reached the RNY code, the elimination of any amino acid at this stage would be strongly selected against and therefore the genetic code was already frozen [35].

There are 20 aaRSs which are divided into two 10-member, non-overlapping classes, I and II, and they provide virtually errorless aminoacylation of tRNAs [36,37]. Therefore, this operational code is non-degenerate [36,37].

In this work, we pose the following question: What are the minimum necessary and sufficient biological and mathematical properties to uniquely determine the primeval RNY code and the SGC?

2: Mathematical Model of the RNY Code

The RNY code consists of codons where the first base is a purine (R), the third is a pyrimidine (Y) and the second is any of them (Table 1). In this code, the wobble position is strictly present on the third base of the triplet. The number of possible RNY codes is $2 \times 4 \times 4 = 32$.

Table 1. RNY code. Amino acids that pertain to class I are in red, and those that correspond to class II are in black.

Amino Acid	Codons	Amino Acid	Codons
Asn	AAC, AAU	Thr	ACC, ACU
Asp	GAC, GAU	Ala	GCC, GCU, GCA, GCG
Ser	AGC, AGU	Ile	AUC, AUU, AUA
Gly	GGC, GGU	Val	GUC, GUU, GUA, GUG

The SGC has been represented in a six-dimensional hypercube [30,38]. Observing that 64 is equal not only to 4^3 but also to 2^6 , the codon table can be organized as a six-dimensional hypercube [30]. In such a model, the set of codons are treated as the 64 vertices of the hypercube, and they are joined by edges which connect codons that differ by a single nucleotide. Each dimension describes a type of mutation, a transition or a transversion acting on each of three bases of any codon. Consequently, we obtain the six dimensions.

This symmetrical model [38] can be partitioned into two classes of amino acids in six dimensions: it displays symmetry groups when the polar and non-polar is used, and the SGC can be broken down into a product of simple groups, reflecting the pattern of degeneracy observed, and the salient fact that substitution did not create it via evolutionary footprints. The model and the Red and Black model [38] have one and the same and the same [37].

Similarly, the RNY code can be represented in a four-dimensional hypercube (Figure 1). This hypercube will be employed to reduce the possible number of mappings by considering its topology and neighborhood properties that codons that codify the same amino acid are neighbors. Note from Figure 1 that the codons for the same amino acid are next to each other due to the fact that they differ in only the third base and therefore they are at a distance of one. A detailed description of the hypercube representing the SGC can be found in Reference [30].

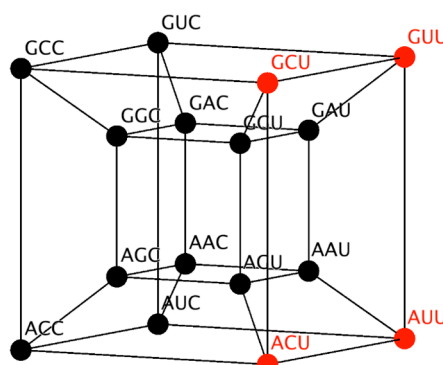


Figure 1. Four-dimensional hypercube that represents the RNY code. Codons for amino acids of class I are in red and those for class II are in black.

3. Combinatorics of the RNY Code

We have noted above that the number of possible codes composed by eight amino acids and 16 triplets is $8^{16} = 2.81 \times 10^{14}$. This number includes codes completely redundant (all codons assigned

6 Combinatorics of the RNY Code

which all amino acids share the same degeneration, as in the present RNY code. Also, there may not be restrictions between the two classes of aaRS and their corresponding amino acids. First, we consider the restriction in which all amino acids are coded by two triplets, and such codes are given by the multinomial coefficient $\frac{16!}{(2, 2, 2, 2, 2, 2, 2, 2)!} = 8! = 17,170$. The present RNY code arranges the triplets so that two codons for the same amino acid are neighbors in the four-dimensional cube. With such a restriction, there are $\binom{4}{1} 8! = 161,280$ possible RNY codes, since there are four possible configurations in the four-dimensional hypercube. This neighborhood property preserves the degeneracy irrespective of the particular wobbling nucleotide, not necessarily the third position. The number $8!$ accounts for the fact that all the permutations in the particular wobbling nucleotide, not necessarily the third position. The number $8!$ accounts for the fact that all the permutations in the assignment of amino acids maintain the property that the two codons that encode the same amino acid must be neighbors.

Considering the third base as the source of variability in the code, the number of possibilities is reduced to $8! = 40,320$. If we consider only the first two bases that determine the amino acid, it is possible to reduce the four-dimensional cube to a three-dimensional cube in which the vertices represent the first two nucleotides (Figure 2a). If the vertices are relabeled to show the codified amino acid, we obtain a phenotypic cube (Figure 2b).

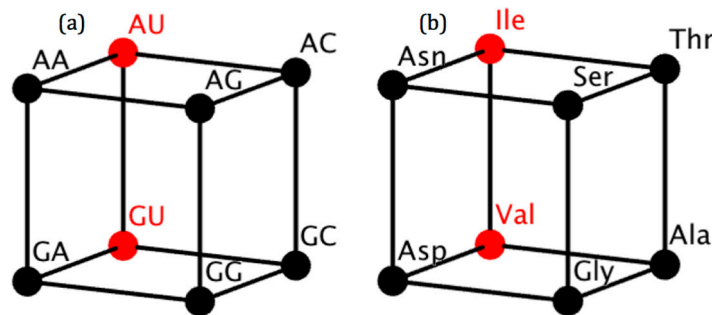


Figure 2. (a) Cube of RNY dinucleotides according to the four-dimensional model of the code. Dinucleotides for class I amino acids are in red; and those for class III are in black; (b) Phenotypic cube of amino acids according to the four-dimensional model of the RNY code. Class I amino acids are in red and those of class II are in black.

If we consider that there are two amino acids that belong to class I and six amino acids that correspond to class III, then there are $2 \binom{8}{2} 6! = 37,440$ possible codes. This calculation comes from

taking two out of the eight amino acids and assigning them to class I, considering its permutations and also the permutations of the amino acids of class II. To maintain the topological properties of the RNY model, four triplets of class I must form a square in the four-dimensional model, or similarly, the dinucleotides must be neighbors, i.e., they are connected by an edge in the cube representation. In this case, there are $2(12)6! = 17,280$ different codes that preserve the aaRSs distribution in the code and in the model. This number arises from the 12 edges available in the cube to join class I amino acids and the permutations of class II amino acids.

In order to maintain the topological properties of the three-dimensional cube, the amino acids of a code must share the neighboring properties of the current RNY code. In other words, if two amino acids are next to each other in the current model, then they are also adjacent in a model constructed by such a code. This property is manifested by the fact that such codes are built by the symmetries of the present model, so that there are 48 different codes that keep the topology of the current code intact.

such a code. This property is manifested by the fact that such codes are built by the symmetries of the present model, so that there are 48 different codes that keep the topology of the current code intact.

The occurrence of glycine as the first amino acid and its assignment to the triplets GGC and GGU as a fixed starting point in the evolution of the SGC impose another restriction, particularly when contrasted with the topology of the four- and three-dimensional cubes, since it fixates isoleucine to AUC and AUU in order to keep the adjacency properties. In this case, there are as many as $\binom{3}{1}^2 = 6$ possible codes, due to the fact that there are three possible positions for valine that maintain its adjacency to isoleucine, and there are two symmetrical configurations (given by a reflection) that maintain the rest of the topology.

In the actual code, all triplets where the middle base is uracile codify for amino acids of class I, and this pattern forces the triplets of valine to be GUC and GUU, which in turn also fixes AGC and AGU for serine. This results in two possible RNY codes, which here and further on will be denoted by \circ RNY and \emptyset RNY. The \circ RNY denotes the actual and original RNY code, whereas \emptyset RNY represents an alternative code in which the codons for threonine and alanine are simultaneously interchanged with the ones of aspartic acid and asparagine, respectively. The fixation of another amino acid would completely constraint the number of RNY possible codes to only one!

4. Evolution of the RNY Code by Means of Frame-Shifts and Transversions

Two genetic codes from which the primeval RNA code could have originated the SGC were derived [14–16]. The primeval RNA code consists of 16 codons that specify eight amino acids (then this code shows a slight degeneration). The extended RNA code type I consists of all codons of the RNY type plus codons obtained by considering the RNA code, but in the second (NYR-type) and third (YRN-type) reading frames. The extended RNA code type II comprises all codons of the RNY type plus codons that arise from transversions of the RNA code in the first (YNY type) and third (RNR) nucleotide bases. Then, by allowing frame-reading mistranslations, we arrived at 48 codons that specify 17 amino acids and the three stop codons. If transversions in the first or third nucleotide bases of the RNY pattern are permitted, then there are also 48 codons that encode for 18 amino acids but no stop codons.

In the context of the frozen concept, it was concluded that considering the symmetries of both extended RNA codes, the primeval RNY code was already frozen and it evolved like a replicating and growing icicle [14]. The composition of both extended codes eventually leads to the actual SGC.

As the RNY is described mathematically as a four-dimensional cube, each extended code comprises a duplication of the RNY cube in order to determine a five-dimensional prism as an intermediate step towards the final six-dimensional cube for the SGC. Supposing one of the two alternative RNY codes as the initial code, the number of possible extended codes can be calculated. Then, assuming, as before, that wobbling occurs principally at the third base, the current degeneration of the code and the topology given by the mathematical model shall be maintained.

If the \circ RNY is used as a cornerstone for the formation of the genetic code, then, regardless of the evolutionary path chosen, there are two SGCs which are compatible with all the assumptions. These are the actual SGC and a second one in which the codifications of AUG and UGG are interchanged with the ones of AUA and UGA, respectively. These modifications make it so that methionine is codified by AUA and tryptophane by UGA, while AUG codes for isoleucine and UGG is a stop signal. The rest of the code remains unaltered.

On the other hand, if \emptyset RNY is used as an initial condition, then there are no possible codes on any evolutionary path which meet all hypotheses. In other words, it is not possible to derive the SGC from \emptyset RNY without violating at least one of the considered properties. This is due to the fact that the mathematical model forbids the possible extended codes that would keep biological properties such as wobbling and the binary division of aaRSs.

5. Discussion

It is possible to gradually add properties to the RNY code to reduce the number of possible codes from 2.81×10^4 to only one. This is done when considering the current properties of degeneracy of the RNY code and the wobble, the aaRSs distribution in the RNY and in the SGC, and finally the mathematical model to represent the genetic code and its induced property of adjacency. The mathematical model plays an important role in the reduction of the possible number of codes. The 37,440 possible RNY codes were obtained by considering the degeneration in the third base and by assuming that the distribution of aaRRs classes is the same as in the current RNY code. Further reductions, up to one code, were only accomplished by the use of our mathematical model. Both evolutionary paths majorly reduce the number of possible genetic codes from the staggering number of 4.18×10^{84} to only two, which consists of the current code and an alternative code with a subtle modification. The alternative RNY code, \emptyset RNY, cannot lead to an SGC that is compatible with all the hypotheses by means of the transversions and frame-shift reading mistranslations. Hence, the SGC evolved from the \circ RNY code.

Novozhilov et al. [39] found that the SGC is a suboptimal random code in regard to robustness to error of translations. Thus, the SGC appears to be a point on an evolutionary trajectory from a random code about halfway to the summit (or to the valley) of the local peak in a rugged fitness landscape.

So far, all we know is terrestrial biology. If life is to be found somewhere else in the universe, and even if its ancestry can be traced back to primitive organisms, the rules of the assignments of codons to amino acids may not necessarily be the same and the amino acids may be even chemically different to those found in known terrestrial life. Different environments and different evolutionary paths on different worlds could result in completely different genetic codes and patterns of evolution.

In conclusion, the SGC is certainly ubiquitous in Earth, and what we would expect to find in living beings on other planets is, precisely, this universal biological property: a genetic coding system.

Acknowledgments: Gabriel S. Zamudio is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and a fellowship recipient from Consejo Nacional de Ciencia y Tecnología (CONACYT) (number: 737920); Marco V. José was financially supported by PAPIIT-IN224015, UNAM, México.

Author Contributions: Gabriel S. Zamudio performed the calculations and figures, wrote a draft of the manuscript; Gabriel S. Zamudio and Marco V. José conceived the work, contributed to ideas, performed the analyses; Marco V. José wrote the manuscript, and prepared the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Crick, F.H.C. The origin of the genetic code. *J. Mol. Biol.* **1968**, *38*, 367–379. [[CrossRef](#)]
2. Woese, C. *The Genetic Code*; Harper and Row: New York, NY, USA, 1967; Chapter 7.
3. Kenneth, D.J.; Ellington, A.D. The search for missing links between self-replicating nucleic acids and the RNA world. *Orig. Life Evol. Biosph.* **1995**, *25*, 515–530.
4. Gilbert, W. The RNA World. *Nature* **1986**, *319*, 618. [[CrossRef](#)]
5. Gesteland, R.F.; Cech, T.R.; Atkins, J.F. *The RNA World*; Cold Spring Harbor Laboratory Press: New York, NY, USA, 1999.
6. Joyce, G.F. The antiquity of RNA-based evolution. *Nature* **2002**, *418*, 214–221. [[CrossRef](#)] [[PubMed](#)]
7. Eigen, M.; Lindemann, B.F.; Tietze, M.; Winkler-Oswatitsch, R.; Dress, A.; Haeseler, A. How old is the genetic code? Statistical geometry of tRNA provides an answer. *Science* **1968**, *244*, 673–679. [[CrossRef](#)]
8. Crick, F.H.C. *On Degenerate Templates and Adaptor Hypothesis Draft*; CSHL Archives Repository: Long Island, NY, USA, 1955.
9. Crick, F.H.C. *On Degenerate Templates and the Adaptor Hypothesis: A Note for the RNA Tie Club*; unpublished but cited by M B Hoagland (1960). In *The Nucleic Acids*; Chargaff, E., Davidson, J.N., Eds.; Academic Press: New York, NY, USA, 1955; Volume 3, p. 349.
10. Crick, F.H.C. On protein synthesis. *Symp. Soc. Exp. Biol.* **1958**, *12*, 138–163. [[PubMed](#)]

11. Crick, F.H.C.; Brenner, S.; Klug, A.; Pieczonik, G. A speculation on the origin of protein synthesis. *Orig. Life* **1976**, *7*, 389–397. [[CrossRef](#)] [[PubMed](#)]
12. Freeland, S.J.; Hurst, L.D. The genetic code is one in a million. *J. Mol. Evol.* **1998**, *47*, 238–248. [[CrossRef](#)] [[PubMed](#)]
13. Rodin, S.N.; Rodin, S.A. Partitioning of aminoacyl-tRNA synthetases in two classes could have been encoded in a strand-symmetric RNA World. *DNA Cell Biol.* **2006**, *25*, 617–626. [[CrossRef](#)] [[PubMed](#)]
14. José, M.V.; Morgado, E.R.; Govezensky, T. An extended RNA code and its relationship to the standard genetic code: An algebraic and geometrical approach. *Bull. Math. Biol.* **2007**, *69*, 215–243. [[CrossRef](#)] [[PubMed](#)]
15. José, M.V.; Morgado, E.R.; Guimarães, R.C.; Zamudio, G.S.; Fariás, S.T.; Bobadilla, J.R.; Sosa, D. Three-dimensional algebraic models of the tRNA code and the 12 graphs for representing the amino acids. *Life* **2014**, *4*, 341–373. [[CrossRef](#)] [[PubMed](#)]
16. José, M.V.; Govezensky, T.; García, J.A.; Bobadilla, J.R. On the evolution of the standard genetic code: Vestiges of scale invariance from the RNA World in current prokaryote genomes. *PLoS ONE* **2009**, *4*, e4340. [[CrossRef](#)] [[PubMed](#)]
17. Wong, J.T. Evolution of the genetic code. *Microbiol. Sci.* **1988**, *5*, 174–181. [[PubMed](#)]
18. Wong, J.T. Coevolution theory of the genetic code at age thirty. *BioEssays* **2005**, *27*, 416–425. [[CrossRef](#)] [[PubMed](#)]
19. Bandhu, A.V.; Aggarwal, N.; Sengupta, S. Revisiting the physico-chemical hypothesis of code origin: An analysis based on code-sequence coevolution in a finite population. *Orig. Life Evol. Biosph.* **2013**, *43*, 465–489. [[CrossRef](#)] [[PubMed](#)]
20. Di Giulio, M. The origin of the genetic code: Matter of metabolism or physicochemical determinism? *J. Mol. Evol.* **2013**, *77*, 131–133. [[CrossRef](#)] [[PubMed](#)]
21. Rouch, D.A. Evolution of the first genetic cells and the universal genetic code: A hypothesis based on macromolecular coevolution of RNA and proteins. *J. Theor. Biol.* **2014**, *357*, 220–244. [[CrossRef](#)] [[PubMed](#)]
22. Miller, S.L. A production of amino acids under possible primitive earth conditions. *Science* **1953**, *15*, 528–529. [[CrossRef](#)]
23. Parker, E.T.; Zhou, M.; Burton, A.S.; Glavin, D.P.; Dworkin, J.P.; Krishnamurthy, R.; Fernández, F.M.; Bada, J.L. A plausible simultaneous synthesis of amino acids and simple peptides on the primordial Earth. *Angew. Chem. Int. Ed. Engl.* **2014**, *28*, 8270–8274. [[CrossRef](#)]
24. Bada, J.L. New insights into prebiotic chemistry from Stanley Miller's spark discharge experiments. *Chem. Soc. Rev.* **2013**, *7*, 2186–2196. [[CrossRef](#)] [[PubMed](#)]
25. Callahan, M.P.; Martin, M.G.; Burton, A.S.; Glavin, D.P.; Dworkin, J. Amino acid analysis in micrograms of meteorite sample by nanoliquid chromatography-high-resolution mass spectrometry. *J. Chromatogr. A* **2014**, *1332*, 30–34. [[CrossRef](#)] [[PubMed](#)]
26. Bernhardt, H.S.; Patrick, W.M. Genetic code evolution started with the incorporation of glycine, followed by other small hydrophilic amino acids. *J. Mol. Evol.* **2014**, *78*, 307–309. [[CrossRef](#)] [[PubMed](#)]
27. Tamura, K. Beyond the Frozen Accident: Glycine Assignment in the Genetic Code. *J. Mol. Evol.* **2015**, *81*, 69–71. [[CrossRef](#)] [[PubMed](#)]
28. Bernhardt, H.S.; Tate, W.P. Evidence from glycine transfer RNA of a frozen accident at the dawn of the genetic code. *Biol. Direct* **2008**, *3*. [[CrossRef](#)] [[PubMed](#)]
29. Parker, E.T.; Cleaves, H.J.; Dworkin, J.P.; Glavin, D.P.; Callahan, M.; Aubrey, A.; Lazcano, A.; Bada, J.L. Primordial synthesis of amines and amino acids in a 1958 Miller H₂S-rich spark discharge experiment. *Proc. Natl. Acad. Sci. USA* **2011**, *5*, 5526–5531. [[CrossRef](#)] [[PubMed](#)]
30. José, M.V.; Morgado, E.R.; Sánchez, R.; Govezensky, T. The 24 possible algebraic representations of the standard genetic code in six and three dimensions. *Adv. Stud. Biol.* **2012**, *4*, 119–152.
31. José, M.V.; Morgado, E.R.; Govezensky, T. Genetic hotels for the standard genetic code: Evolutionary analysis based upon novel three-dimensional algebraic models. *Bull. Math. Biol.* **2011**, *73*, 1443–1476. [[CrossRef](#)] [[PubMed](#)]
32. Eigen, M.; Winkler-Oswatitsch, R. Transfer-RNA: An early gene? *Naturwissenschaften* **1981**, *68*, 282–292. [[CrossRef](#)] [[PubMed](#)]
33. Miller, S.L.; Urey, H.C.; Oró, J. Origin of organic compounds on the primitive earth and in meteorites. *J. Mol. Evol.* **1976**, *9*, 59–72. [[CrossRef](#)] [[PubMed](#)]

34. Trifonov, E.N. Consensus temporal order of amino acids and evolution of the triplet code. *Gene* **2000**, *261*, 139–151. [[CrossRef](#)]
35. José, M.V.; Zamudio, G.S.; Palacios-Pérez, M.; Bobadilla, J.R.; Farías, S.T. Symmetrical and thermodynamic properties of phenotypic graphs of amino acids encoded by the primeval RNY code. *Orig. Life Evol. Biosph.* **2015**, *45*, 77–83. [[CrossRef](#)] [[PubMed](#)]
36. de Pouplana, L.R.; Schimmel, P. Aminoacyl-tRNA synthetases: Potential markers of genetic code development. *Trends Biochem. Sci.* **2001**, *26*, 591–596. [[CrossRef](#)]
37. Schimmel, P.; Giégé, R.; Moras, D.; Yokoyama, S. An operational RNA code for amino acids and possible relationship to genetic code. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 8763–8768. [[CrossRef](#)] [[PubMed](#)]
38. José, M.V.; Zamudio, G.S.; Morgado, E.R. A unified model of the standard genetic code. *R. Soc. Open Sci.* **2017**, *4*, 160908. [[CrossRef](#)]
39. Novozhilov, A.S.; Wolf, Y.I.; Koonin, E. Evolution of the genetic code: Partial optimization of a random code for robustness to translation error in a rugged fitness landscape. *Biol. Direct* **2007**, *2*, 1–24. [[CrossRef](#)] [[PubMed](#)]



© 2017 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Apéndice 3

Phenotypic Graphs and Evolution Unfold the Standard Genetic Code as the Optimal

Gabriel S. Zamudio¹ · Marco V. José¹

Received: 2 July 2017 / Accepted: 16 October 2017 /

Published online: 29 October 2017

© Springer Science+Business Media B.V. 2017

Abstract In this work, we explicitly consider the evolution of the Standard Genetic Code (SGC) by assuming two evolutionary stages, to wit, the primeval RNY code and two intermediate codes in between. We used network theory and graph theory to measure the connectivity of each phenotypic graph. The connectivity values are compared to the values of the codes under different randomization scenarios. An error-correcting optimal code is one in which the algebraic connectivity is minimized. We show that the SGC is optimal in regard to its robustness and error-tolerance when compared to all random codes under different assumptions.

Keywords Standard genetic code · Graph theory · Network theory · Evolution genetic code · Error-tolerance

Introduction

The standard genetic code (SGC) is almost universal. This feature supports the hypothesis of the existence of Last Common Ancestor Universal (LUCA) and it led to the proposal of Crick's "Frozen accident hypothesis" (Crick 1968). This hypothesis states that the SGC does not change and was fixed by an accident. Arguing that changes in the SGC would be lethal or be strongly selected against, Crick stated the most accepted hypothesis for the evolution of the genetic code (Crick 1968). Different proposals for the origin and evolution of the genetic code have been made (Woese 1973; Rodin et al. 2011; Carter and Wolfenden 2015; Ribas de Pouplana and Schimmel 2001; Delarue 2007; Rodin and Rodin 2008; José et al. 2017; Wong 1988, 2005; Di Giulio 2013; Bandhu et al. 2013; Rouch 2014).

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11084-017-9552-3>) contains supplementary material, which is available to authorized users.

✉ Marco V. José
marcojose@biomedicas.unam.mx

¹ Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, C.P. 04510 Ciudad de México CDMX, Mexico

The SGC is a dictionary of three letter words based on the alphabet (A, U, C, G). The words are known as codons. Sixty-one codons codify for 20 amino acids and three codify the stop signals. Glycine is considered the first amino acid incorporated into the genetic code (Bernhardt and Patrick 2014; Tamura 2015). The SGC is commonly represented in a table (Table 1), with the codons written in the 5' to 3' direction, and it allows us to read off this code directly (Crick 1968). One property that immediately stands out from this representation is its degeneracy in the third base of some codons. This block structure accounts for the Crick's

Table 1 The table of the SGC

		Second Letter				
		U	C	A	G	
First Letter	U	UUU	UCU	UAU	UGU	U
		Phe	Ser	Tyr	Cys	C
		UUC		UAC	UGC	A
		UUA	UCA	UAA	UGA	Stop
	Leu	UCG	UAG	UGG	Tyr	
	G					
	C	CUU	CCU	CAU	CGU	U
		Leu	Pro	His	Arg	C
				CUC		CAC
		CUA	CCA	CAA	CGA	G
	CUG	CCG	CAG	CGG		
	A	AUU	ACU	AAU	AGU	U
		Ile	Thr	Asn	Ser	C
				AUC	ACC	AGC
		AUA	ACA	AAA	AGA	Arg
Met	ACG	AAG	AGG	G		
G	GUU	GCU	GAU	GGU	U	
	Val	Ala	Asp	Gly	C	
			GUC		GCC	GGC
	GUA	GCA	GAA	GGA	G	
GUG	GCA	GAG	GGG			

wobble hypothesis (Crick 1958; Crick et al. 1976). The wobble hypothesis states that the third base is much less specific than the other two, and so, it is said to wobble and allow similar codons to codify for the same amino acid. The SGC is shown in Table 1, where the first and second bases of the codons arrange the amino acids in blocks, whereas the wobble property resides in the third base.

The RNY (purine-any base-pyrimidine) subcode is mostly considered as the primeval genetic code (Eigen and Schuster 1978; Eigen and Winkler-Oswatitsch 1981). It is composed by 16 codons and 8 amino acids, with two codons for each amino acid. Most of these amino acids were also found in Miller's experiments and observations on abundant amino acids and in meteorites (Miller 1953; Miller et al. 1976). The SGC has been theoretically derived from the RNY code through a process of symmetry breakings (José et al. 2007, 2014) that were identified with two evolutionary pathways. The first path is a degenerate RNA code which can be translated in the 1st (RNY), 2nd (NYR), and the 3rd (YRN) reading frames, and the second path is obtained by transversions in the 1st (YNY) and 3rd (RNR) nucleotide bases of the 16 codons of the RNA subcode. The composition of both intermediate subcodes led straightforward to the SGC using the RNY as primeval code. The SGC was mathematically modeled in a hypercube of six dimensions (6D) where the vertices of the cube represent the codons and the edges join codons that differ in a single nucleotide (José et al. 2017). In the 6D-hypercube, the RNY code is represented by a 4-dimensional hypercube (Zamudio and José 2017) and the evolutionary steps are expansions to higher dimensions in order to finally reach the 6D-hypercube of the SGC. Phenotypic graph representations of amino acids based on the topology of the SGC hypercube has been developed (José et al. 2015). A phenotypic graph is one in which the vertices represent the 21 signals of the SGC and the edges join amino acids or the stop signal based on the adjacencies of the SGC hypercube (José et al. 2014, 2015; Zamudio and José 2017).

The current block structure of the genetic code has also been analyzed to understand the processes involved in the assignment of amino acids to codons (Woese et al. 1966; Caporaso et al. 2005). The robustness of the SGC and its capacity to tolerate errors in translation by misreading of the codons have been analysed (Alff-Steinberger 1969; Ardell and Sella 2002). This feature was analysed with random codes (Novozhilov et al. 2007). Stochastic simulations showed that the SGC is not optimal for error-minimization but when other biological properties were taken into account it was found to be suboptimal in a fitness landscape (Wong 1980; Haig and Hurst 1991; Freeland et al. 2000; Novozhilov et al. 2007).

In this work, we usher in both network and graph theory as a novel approach to examine the evolution of the genetic code. We analyze the connectivity properties of the phenotypic graphs that are extracted from the genetic codes that arise from two evolutionary pathways assuming RNY as a primeval code. The measure known as algebraic connectivity reflects the general connectivity of a network (de Abreu 2007; Newman 2010). As this measure gets higher, a network is more connected. The application of this measure to the phenotypic graphs will determine the connectivity of the 21 signals of the SGC and reflect the associations of the codons in the 6D-hypercube model. These connectivity values are compared to the values of the codes under different randomization scenarios. The codon – amino acid associations are fixed in every stage of the evolution of the genetic code before proceeding to the next evolutionary step.

We show that the SGC is indeed optimal by analyzing the algebraic connectivity of the phenotypic graphs of the SGC, the extended codes (intermediate evolutionary steps), and the RNY code (primeval code).

Methods

Graph theory is a branch of discrete mathematics devoted to the analysis of graphs. A graph is an object composed by vertices and edges. The vertices represent objects of any kind and the edges reflect some relation between those objects (Newman 2010). Several measures have been developed to describe the connectivity of a graph. Most of them assign to each vertex a centrality or connectivity score (Newman 2010). The algebraic connectivity is a connectivity measure for a set of vertices or for the whole graph. This measure is denoted by $\alpha(G)$, where G stands for a graph. It is given by the second smallest eigenvalue of the Laplacian matrix of G (de Abreu 2007; Newman 2010) and represents a parameter to measure the connectivity of a graph. The adjacency matrix is important in graph theory because it captures the entire structure of a network and whose matrix properties are used to characterize the properties of edges and vertices. There is another matrix, the Laplacian matrix closely related to the adjacency matrix that can also tell us much about the network structure. In fact, the Laplacian matrix of a graph G is given by, the adjacency matrix of G minus the identity matrix. The Laplacian is a symmetric matrix, and so has real eigenvalues. All the eigenvalues of the Laplacian are also non-negative. While the eigenvalues of the Laplacian cannot be negative, they can be zero, and in fact the Laplacian always has at least one zero eigenvalue. Since there are no negative eigenvalues, this is the lowest of the eigenvalues of the Laplacian. By convention, the eigenvalues are numbered in ascending order: $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$. So we always have $\lambda_1 = 0$. The second eigenvalue of the graph Laplacian λ_2 is non-zero if and only if the network is connected. The algebraic connectivity is widely used in algorithms aimed to expand graphs, and the eigenvector associated to the algebraic connectivity, known as Fiedler vector, is used for combinatorial optimization problems (de Abreu 2007).

Inhere, the evolutionary pathways of the SGC with a RNY subcode as its ancestor is considered. The genetic codes from each stage of the evolution path are subject to different degrees of randomization and set in the 6D-dimensional representation of the SGC, in which the vertices of a 6D-hypercube represent the codons, and the edges join codons that differ by a single nucleotide (José et al. 2017). Then, its phenotypic graphs were calculated. The phenotypic graph represents the phenotypic expression of the codon hypercube; the vertices represent the 20 amino acids and the stop signal. In the phenotypic graph, two signals (amino acid or stop signal) of the genetic code are joined if there exist two codons for those signals that are adjacent in the 6D-model (José et al. 2014, 2015; Zamudio and José 2017). A lower algebraic connectivity in the phenotypic graphs reflects a general absence of edges joining the amino acids in the phenotypic graph. In turn, the latter reflects that the codons codifying for the same amino acid are joined, and therefore, they are similar, instead of being randomly scattered across the hypercube. An error-correcting optimal code is one in which the algebraic connectivity is minimized. If the amino acids were randomly associated to the codons, the block structure of the genetic code vanishes and the resulting phenotypic graph would present a high algebraic connectivity.

Starting from the RNY subcode, two evolutionary pathways have been proposed (José et al. 2009, 2014). One path comprehends an extension of the RNY subcode by means of frame-shift reading mistranslations; this generates the codes NYR and YRN in addition to the original RNY. This path is known as the Extended code type I (Ex1). The second path is reached by transversions in the first and the third base of the RNY subcode; this code comprises the codons of the form RNR and YNY. The second path is known as Extended code type II (Ex2). By complementing both Extended codes, the rest of the codons are generated and the SGC is completed to 64 codons.

For the RNY subcode, random codes in which the amino acids were randomly permuted and assigned to the 16 RNY codons were calculated. RNYp denotes these random codes. Three levels of randomization were calculated for each of the extended codes. Ex1s and Ex2s denote the codes that resulted from a random permutation of the amino acids and stop signal of the SGC and then the extended codes were extracted. Ex1p and Ex2p denote the codes that arise from a restricted permutation of the signals present in each of the extended codes. Finally, Ex1d and Ex2d denote the extended codes with a random degeneracy of the signals present on each code. For the complete genetic code, three distinct levels of randomization were also considered. Codes with random degeneracy of the 21 signals are denoted by SGCr. Codes with the amino acids and the stop signal randomly permuted are SGCp. The third level is denoted by SGCrd in which all the properties of the wobble and the degeneracy are preserved; the wobbling is present in the third base. The SGCrd codes represents codes that only permute blocks that maintain the third wobbling base of the SGC. For each random code, 5000 permutations were calculated and the algebraic connectivity for each phenotypic graph obtained by each random code case was computed.

Results

For all the randomized codes, the range of values or intervals between the minimum and maximum of the algebraic connectivity is presented in Table 2. The algebraic connectivity of the RNY, Ex1, Ex2 and SGC are also included. For all the random control codes of the Extended codes, none of them presented a minimum of algebraic connectivity that falls below the algebraic connectivity of both Ex1 and Ex2. There were only 3 codes out of the 5000 random codes of the RNY code, that presented the same connectivity as the actual RNY whose connectivity is equal to 2. These three codes (RNYp) differ from the actual RNY by a misplacement of Glycine in the triplets GGC and GGU (Table 3).

In the Ex2s randomization, 4 out of the 5000 permutations presented a lower connectivity than the actual Extended code type II. These permutations displayed alterations in the RNY code. Recall that the random controls of the SGC were as follows: pure random degeneracy (SGCr), fixed degeneracy without wobbling (SGCp), and fixing wobbling but shuffling the codon-amino acids assignments (SGCrd). Note in Table 2 that the maximum values of the

Table 2 The intervals between the minimum and maximum of algebraic connectivity calculated for each randomized code under different randomization hypotheses

The algebraic connectivity of the actual codes, at various stages in evolution, are shown in the first lines

Code	Randomized code	Interval
RNY Code	RNY	2
	RNYp	(2–6)
Extended Code Type I	Ex1	1.331
	Ex1s	(1.527–4.936)
	Ex1d	(1.699–6.249)
	Ex1p	(1.822–4.704)
Extended Code Type II	Ex2	1.733
	Ex2s	(1.658–6.097)
	Ex2d	(1.763–6.355)
	Ex2p	(1.821–5.634)
Standard Genetic Code	SGC	2.049
	SGCr	(2.392–8.306)
	SGCp	(1.807–5.867)
	SGCrd	(1.877–2.170)

Table 3 RNYp codes with the same algebraic connectivity as the RNY code

Code	RNY Code	RNYp Code1	RNYp Code 2	RNYp Code 3
AAC	Asn	Asp	Asn	Asn
AAU	Asn	Asp	Val	Asn
ACC	Thr	Ser	Thr	Ser
ACU	Thr	Thr	Ile	Asp
AGC	Ser	Thr	Val	Ala
AGU	Ser	Asn	Ser	Ala
AUC	Ile	Ser	Asp	Gly
AUU	Ile	Asn	Thr	Thr
GAC	Asp	Ile	Ala	Gly
GAU	Asp	Ile	Gly	Thr
GCC	Ala	Val	Ile	Val
GCU	Ala	Gly	Asn	Ile
GGC	Gly	Gly	Ser	Ser
GGU	Gly	Ala	Ala	Asp
GUC	Val	Val	Gly	Ile
GUU	Val	Ala	Asp	Val
Misplacements are found in the GGC and GGU codons	Algebraic Connectivity	2	2	2

connectivity of the random controls of the SGC, decrease as more restrictions to the controls are added, reflecting that codons codifying for the same amino acid tend to be closer together in the 6D-hypercube. The hexacodonic amino acids Serine, Leucine and Arginine, present two codons that are less similar than the other four in which the third base wobble, especially Serine. In other words, in hexacodonic amino acids, if the 6 codons are more similar among them (to keep the second base invariant), then the connectivity of the phenotypic graph will decrease. Some random codes of the SGC present a lower connectivity than the current SGC. The SGCp random control showed one code and the SGCrd presented 1825 codes with lower connectivity. All of them presented modifications in the RNY subcode and in the Extended codes. Details of the codes with lower algebraic connectivity are presented in [Appendix](#).

Discussion

We show that the SGC is indeed optimal when its evolutionary stages from the primeval RNY code, and the extended codes are considered. This result was achieved by calculating the connectivity properties of the phenotypic graphs of each code. The different degrees of randomization applied to the SGC and the Extended codes allowed measuring the effect of different properties of the genetic code to assign the same amino acid to similar codons. The RNY subcode and the Extended codes delineate concrete stages to analyze the robustness and error correction properties of the genetic code, in contrast to previous approaches which ignored the evolution of the SGC (Haig and Hurst 1991; Novozhilov et al. 2007). The randomized codes did show that more optimal codes exist as found in other works (Wong 1980; Haig and Hurst 1991; Freeland et al. 2000; Novozhilov et al. 2007). However, the codes that resulted with a lower connectivity than the present SGC exhibited modifications on the initial stages of the evolutionary pathway, i.e. the amino acids associated to codons of the Extended codes or the RNY subcode are different from the ones currently present in those codes. In a general perspective the codes with lower connectivity would be more resistant to

errors, yet their occurrence would require major codon swaps to achieve those codes and would not be evolutionary optimal as proposed by Di Giulio (1989). Codon swaps for the reassignment of amino acids to triplets has been proposed as one of the principal mechanisms for the evolution of the genetic code (Szathmáry 1991). In between the evolutionary stages of the SGC, derived by the symmetry-breaking model, codon swaps must have presumably occurred in order to produce the present assignments of those codons. Once any of the two Extended codes were reached it was fixed and no amino acid reassignments further occurred; the same fixation is extended to the core RNY subcode. The scaling properties of the distance series of each codon of the RNY, Extended codes, and the SGC show critical scale invariance and this property is a universal vestige in genomes of Eubacteria and Archea (José et al. 2009).

The 6D-model of the genetic code (José et al. 2007) has been shown to be equivalent to the Rodin-Ohno model (Rodin and Ohno 1995) of the genetic code (José et al. 2017). It shows symmetries relative to the Woese's polar requirement scales of amino acids (Woese et al. 1966), and to the partition of the code derived by the class of aminoacyl-tRNA synthetases (aaRSs) associated to the amino acids (José et al. 2017). When restricted to the RNY subcode, its phenotypic graph has been coupled with the division of aaRS in order to derive the biological properties that uniquely identify the present SGC (Zamudio and José 2017). The regularity of this core code is reflected in its phenotypic graphs (José et al. 2015) when a SGC model in lower dimensions is considered (José et al. 2012).

de Fariás et al. (2014) proposed an origin for the coding system based on the co-evolution of tRNAs and aaRS, and further driven by changes in the second base of the anticodon that affected its hydrophathy. Other mechanisms driving the evolution of the genetic code include the polarity of amino acids (Di Giulio 1989), and the configuration of peptides formed by the genetic code (Alff-Steinberger 1969). All these pressures fixed and froze the SGC at different stages when the codons of the evolutionary paths were assigned to the amino acids that constitute a genetic code more robust and with high error-tolerance capacity, leading ultimately to the completion SGC.

Acknowledgments Gabriel S. Zamudio is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and a fellowship recipient from Consejo Nacional de Ciencia y Tecnología (CONACYT) (number: 737920). M.V.J. was financially supported by PAPIIT-IN224015; UNAM; México. We thank Juan R. Bobadilla for the technical computer support.

References

- Alff-Steinberger C (1969) The genetic code and error transmission. *Proc Natl Acad Sci U S A* 64:584–591. <https://doi.org/10.1073/pnas.64.2.584>
- Ardell DH, Sella G (2002) No accident: genetic codes freeze in error-correcting patterns of the standard genetic code. *Philos Trans R Soc Lond Ser B Biol Sci* 357:1625–1642. <https://doi.org/10.1098/rstb.2002.1071>
- Bandhu AV, Aggarwal N, Sengupta S (2013) Revisiting the physico-chemical hypothesis of code origin: an analysis based on code-sequence coevolution in a finite population. *Orig Life Evol Biosph* 43:465–489. <https://doi.org/10.1007/s11084-014-9353-x>
- Bernhardt HS, Patrick WM (2014) Genetic code evolution started with the incorporation of glycine, followed by other small hydrophilic amino acids. *J Mol Evol* 78:307–309. <https://doi.org/10.1007/s00239-014-9627-y>
- Caporaso JG, Yarus M, Knight R (2005) Error minimization and coding triplet/binding site associations are independent features of the canonical genetic code. *J Mol Evol* 61:597–607. <https://doi.org/10.1007/s00239-004-0314-2>
- Carter CW, Wolfenden R (2015) tRNA acceptor stem and anticodon bases form independent codes related to protein folding. *Proc Natl Acad Sci* 112:7489–7494. <https://doi.org/10.1073/pnas.1507569112>
- Crick FHC (1958) On protein synthesis. *Symp Soc Exp Biol* 12:138–166

- Crick FHC (1968) The origin of the genetic code. *J Mol Biol* 38:367–379. [https://doi.org/10.1016/0022-2836\(68\)90392-6](https://doi.org/10.1016/0022-2836(68)90392-6)
- Crick FHC, Brenner S, Klug A, Pieczek G (1976) A speculation on the origin of protein synthesis. *Orig Life* 7: 389–397. <https://doi.org/10.1007/BF00927934>
- de Abreu NMM (2007) Old and new results on algebraic connectivity of graphs. *Linear Algebra Appl* 423:53–73. <https://doi.org/10.1016/j.laa.2006.08.017>
- de Fariás ST, do Rêgo TG, José MV (2014) Evolution of transfer RNA and the origin of the translation system. *Front Genet* 5:303
- Delarue M (2007) An asymmetric underlying rule in the assignment of codons: possible clue to a quick early evolution of the genetic code via successive binary choices. *RNA* 13:161–169. <https://doi.org/10.1261/rna.257607>
- Di Giulio M (1989) The extension reached by the minimization of the polarity distances during the evolution of the genetic code. *J Mol Evol* 29:288–293. <https://doi.org/10.1007/BF02103616>
- Di Giulio M (2013) The origin of the genetic code: matter of metabolism or physicochemical determinism? *J Mol Evol* 77:131–133. <https://doi.org/10.1007/s00239-013-9593-9>
- Eigen M, Schuster P (1978) The hypercycle - a principle of natural self-organization part b: the abstract hypercycle. *Naturwissenschaften* 65:7–41. <https://doi.org/10.1007/BF00420631>
- Eigen M, Winkler-Oswatitsch R (1981) Transfer-RNA, an early gene? *Naturwissenschaften* 68:282–292. <https://doi.org/10.1007/BF01047470>
- Freeland SJ, Knight RD, Landweber LF, Hurst LD (2000) Early fixation of an optimal genetic code. *Mol Biol Evol* 17:511–518. <https://doi.org/10.1093/oxfordjournals.molbev.a206331>
- Haig D, Hurst LD (1991) A quantitative measure of error minimization in the genetic-code. *J Mol Evol* 33:412–417. <https://doi.org/10.1007/bf02103132>
- José MV, Morgado ER, Govezensky T (2007) An extended RNA code and its relationship to the standard genetic code: an algebraic and geometrical approach. *Bull Math Biol* 69:215–243. <https://doi.org/10.1007/s11538-006-9119-3>
- José MV, Govezensky T, García JA, Bobadilla JR (2009) On the evolution of the standard genetic code: vestiges of critical scale invariance from the RNA world in current prokaryote genomes. *PLoS One*. <https://doi.org/10.1371/journal.pone.0004340>
- José MV, Morgado ER, Sanchez R, Govezensky T (2012) The 24 possible algebraic representations of the standard genetic code in six or in three dimensions. *Adv Stud Biol* 4:119–152
- José MV, Morgado ER, Guimarães RC et al (2014) Three-dimensional algebraic models of the tRNA code and 12 graphs for representing the amino acids. *Life (Basel, Switzerland)* 4:341–373. <https://doi.org/10.3390/life4030341>
- José MV, Zamudio GS, Palacios-Pérez M et al (2015) Symmetrical and thermodynamic properties of phenotypic graphs of amino acids encoded by the primeval RNY code. *Orig Life Evol Biosph* 45:77–83. <https://doi.org/10.1007/s11084-015-9427-4>
- José MV, Zamudio GS, Morgado ER (2017) A unified model of the standard genetic code. *R Soc Open Sci* 4: 160908. <https://doi.org/10.1098/rsos.160908>
- Miller SL (1953) A production of amino acids under possible primitive earth conditions. *Science* (80-) 117:528–529. <https://doi.org/10.1126/science.117.3046.528>
- Miller SL, Urey HC, Oró J (1976) Origin of organic compounds on the primitive earth and in meteorites. *J Mol Evol* 9:59–72. <https://doi.org/10.1007/BF01796123>
- Newman MEJ (2010) *Networks : an introduction*. Oxford University Press, Oxford
- Novozhilov AS, Wolf YI, Koonin EV (2007) Evolution of the genetic code: partial optimization of a random code for robustness to translation error in a rugged fitness landscape. *Biol Direct* 2:24. <https://doi.org/10.1186/1745-6150-2-24>
- Ribas de Pouplana L, Schimmel P (2001) Aminoacyl-tRNA synthetases: Potential markers of genetic code development. *Trends Biochem Sci* 26:591–596
- Rodin SN, Ohno S (1995) Two types of aminoacyl-trna synthetases could be originally encoded by complementary strands of the same nucleic ACID. *Orig Life Evol Biosph* 25:565–589. <https://doi.org/10.1007/BF01582025>
- Rodin SN, Rodin AS (2008) On the origin of the genetic code: signatures of its primordial complementarity in tRNAs and aminoacyl-tRNA synthetases. *Heredity (Edinb)* 100:341–355. <https://doi.org/10.1038/sj.hdy.6801086>
- Rodin AS, Szathmáry E, Rodin SN (2011) On origin of genetic code and tRNA before translation. *Biol Direct* 6: 14. <https://doi.org/10.1186/1745-6150-6-14>
- Rouch DA (2014) Evolution of the first genetic cells and the universal genetic code: a hypothesis based on macromolecular coevolution of RNA and proteins. *J Theor Biol* 357:220–244. <https://doi.org/10.1016/j.jtbi.2014.06.003>

- Szathmáry E (1991) Codon swapping as a possible evolutionary mechanism. *J Mol Evol* 32:178–182. <https://doi.org/10.1007/BF02515390>
- Tamura K (2015) Beyond the frozen accident: glycine assignment in the genetic code. *J Mol Evol* 81:69–71
- Woese CR (1973) Evolution of the genetic code. *Naturwissenschaften* 60:447–459. <https://doi.org/10.1007/BF00592854>
- Woese CR, Dugre DH, Saxinger WC, Dugre SA (1966) The molecular basis for the genetic code. *Proc Natl Acad Sci U S A* 55:966–974. <https://doi.org/10.1073/pnas.55.4.966>
- Wong JT (1980) Role of minimization of chemical distances between amino acids in the evolution of the genetic code. *Proc Natl Acad Sci U S A* 77:1083–1086
- Wong JT (1988) Evolution of the genetic code. *Microbiol Sci* 5:174–181
- Wong JT-F (2005) Coevolution theory of the genetic code at age thirty. *BioEssays* 27:416–425. <https://doi.org/10.1002/bies.20208>
- Zamudio GS, José MV (2017) On the uniqueness of the standard genetic code. *Life (Basel, Switzerland)* 7:7. <https://doi.org/10.3390/life7010007>.

Origins of Life & Evolution of the Biosphere is a copyright of Springer, 2018. All Rights Reserved.

Apéndice 4

Article

Symmetrical Properties of Graph Representations of Genetic Codes: From Genotype to Phenotype

Marco V. José *  and Gabriel S. Zamudio 

Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, Mexico D.F. 04510, Mexico; gazaso92@gmail.com

* Correspondence: marcojose@biomedicas.unam.mx

Received: 4 August 2018; Accepted: 5 September 2018; Published: 8 September 2018



Abstract: It has long been claimed that the mitochondrial genetic code possesses more symmetries than the Standard Genetic Code (SGC). To test this claim, the symmetrical structure of the SGC is compared with noncanonical genetic codes. We analyzed the symmetries of the graphs of codons and their respective phenotypic graph representation spanned by the RNY (R purines, Y pyrimidines, and N any of them) code, two RNA Extended codes, the SGC, as well as three different mitochondrial genetic codes from yeast, invertebrates, and vertebrates. The symmetry groups of the SGC and their corresponding phenotypic graphs of amino acids expose the evolvability of the SGC. Indeed, the analyzed mitochondrial genetic codes are more symmetrical than the SGC.

Keywords: standard genetic code; mitochondrial codes; phenotypic graphs; graph theory; group theory; evolution

1. Introduction

The discovery of the structure of DNA [1] and the decipherment of the Standard Genetic Code (SGC) [2,3] are landmarks of scientific achievements. The elucidation of the origin and evolution of the SGC is a central problem in evolutionary biology. The SGC is nearly universal, with some minor exceptions. Crick proposed the frozen accident hypothesis to account for the universality of the SGC [4]. The universality of the SGC immediately implied a Last Universal Common Ancestor (LUCA). Therefore, evolution has to do with preserving or fixing some necessary properties of life. Given the astonishing diversity of life in the history of the biosphere, the fact that the SGC is frozen indicates that all organisms are phylogenetically related.

Attempts at thawing the origin and evolution of the frozen SGC have been numerous. Symmetries in the SGC have been analyzed by examining the transfer RNA (tRNA) [5,6], the aminoacyl-tRNA-synthetases (aaRSs) [7–10], and mathematical models searching hidden symmetries [11,12]. The hypercube algebraic representation has allowed the analysis of the evolution of the SGC and a variety of its biological properties. The SGC, as derived from the primeval genetic code [5], and the Rodin–Ohno model [9] are one and the same, that is, these seemingly different models of the genetic code are mathematically equivalent [13]. Hence, the 6D algebraic model unifies different models of the genetic code [13].

The genetic code is a dictionary composed of words three letters long, known as codons or triplets, each letter a nucleotide base. There are four basic nucleotides in the DNA, to wit, adenine (A), cytosine (C), guanine (G), and thymine (T). During the translation process of the DNA, thymine nucleotides are replaced by uracil (U) in the RNA. This constitutes a set of 64 possible codons, which codify for 20 canonical amino acids and a stop signal. The genetic code is degenerated, as more than one codon can codify for a given amino acid. This degeneracy property usually occurs in the third base of a codon and is known as the wobble property [14–16].

The nucleotide bases can be divided according to their chemical properties into: strong $S = \{G, S\}$ and weak $W = \{U, A\}$; amino nucleotides $M = \{C, A\}$, keto nucleotides $K = \{U, G\}$, and nucleotide bases of the same chemical kind: pyrimidines $Y = \{C, U\}$ and purines $R = \{A, G\}$ [17].

Diverse genetic codes occur in a cell, for example, the SGC, the genetic code of mitochondria, the genetic code of chloroplasts, the anticodon code of tRNAs, ribosomal codes.

The mitochondrion is the major energy provider of the Eukaryotic cell [18]. Mitochondria produces ATP by oxidizing the major products of glucose: pyruvate, and NADH [18]. This type of cellular respiration known as aerobic respiration, is dependent on the presence of oxygen. When oxygen is scarce, the glycolytic products will be metabolized by anaerobic fermentation, a process that is independent of the mitochondria [18]. Mitochondria also contribute to many physiological processes, such as calcium homeostasis, apoptosis, lipid and amino acid metabolism [19–21].

The different genetic codes that have been encountered so far (e.g., mitochondrial, *Euplotes*, some ciliate protozoans, *Tetrahymena*) are considered to have evolved from the SGC [22]. Most of the noncanonical codes arise from alterations in the transfer RNA (tRNA) by post-transcriptional modifications, such as base modification or RNA editing, rather than by substitutions within tRNA anticodons. Typically, variations occur in the uncoded amino acids (Met and Trp) and in the stop codons UAG (amber), UGA (opal), and UAA (ochre). However, the freezing of the code is supported by the fact that the 20 natural amino acids have been stringently selected over the course of the evolution (with the notable exception of selenocysteine and pyrrolysine [23]). Then, the SGC can evolve but at a glacial rate.

To examine the symmetries of the SGC, it is necessary to unleash it from the traditional 2D representation of the Table of the Genetic Code [4].

The SGC exhibits an exact symmetry under a Galois Field of 4 elements, also known as the Klein Four-Group (an Abelian (commutative) group of order 4 where each element is its own inverse) [24]. The Klein Four-Group emerges from the primeval RNY code that evolved until the formation of the SGC. This symmetry has been selected since the origin and during the evolution of the genetic code. The SGC has been derived by assuming a primeval genetic code, RNY [25]. This primeval RNY code was composed of 16 codons that codify for 8 amino acids (slight degeneration) and was proposed by Eigen 40 years ago [25]. Two evolutionary paths have been established to reach the SGC from the RNY code [17,26]. These paths consist in permitting frame-shift reading-mistranslations or transversions in the first and third base of the codons. The SGC has been modeled as a six-dimensional (6D) binary hypercube $(\mathbb{Z}_2)^6$, where $\mathbb{Z}_2 = \{0, 1\}$ is the binary field of 2 elements, also known as $GF(2)$ the Galois Field of 2 elements. The binary hypercube is a 2^6 —Klein Group [27]. In the 6D hypercube, the vertices are indexed by the codons [13,26]. The hypercube of codons has been further transformed into its phenotypic graph representation [13,28–30], where the new vertices are the amino acids and the stop signal.

One goal of the present work is to determine if these symmetries were selected since the origin of the primeval code and preserved during its evolution until the formation of the SGC. In this work, the hypothesis that the symmetry groups must allow us to predict the possible symmetry breaking groups to determine the evolvability of the SGC is put forward. To this end, we examine how the SGC has led to new genetic codes by determining their symmetries. We analyze the symmetries of the graphs of codons and their respective phenotypic graph representation spanned by the RNY code, the two RNA Extended codes, and the complete code of 64 codons that comprises the SGC, as well as three different mitochondrial genetic codes from yeast, invertebrates, and vertebrates. In general, the SGC has evolved into more symmetrical mitochondrial codes.

2. Material and Methods

The four nucleotides of the RNA alphabet, A, U, C, and G, can be arranged in three different ways as the vertices of a square that are not symmetrically equivalent, and in one extra way considering the two diagonals of the square (Figure 1).

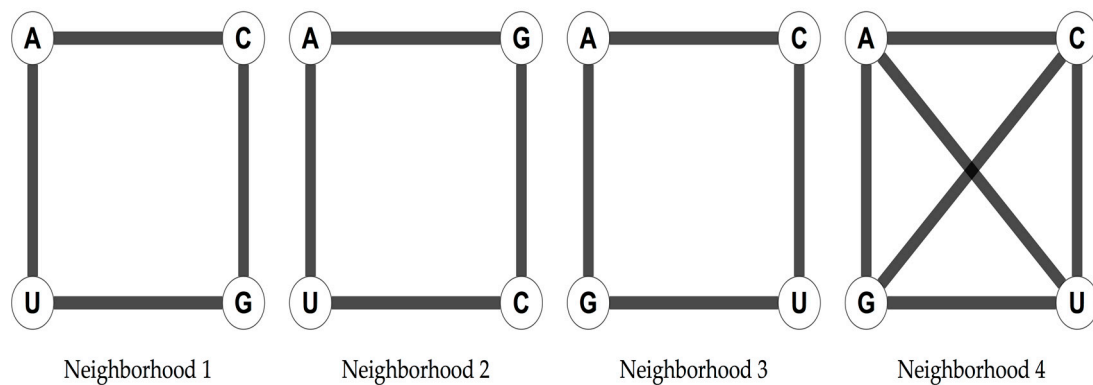


Figure 1. Four possible arrangements of the four nucleotides as the vertices of a square that are not symmetrically equivalent.

The arrangement in a square has been shown to yield a 6D hypercube when considering the 64 possible triplets [13,26]. The genetic code is then represented as a 6D hypercube, which can be interpreted as a graph of vertices representing the codons, and edges joining the codons at distance one, making it possible to analyse its symmetries through the group of automorphisms of the graph [13]. This group consists of all the bijective functions of the graph G , that preserve its adjacencies. These automorphisms comprise all the isometric transformations of the cube. The 6D hypercube arises when the triplets are used as vertices of a graph. Two vertices, or triplets, will be joined by an edge if they differ by one letter, and the different letters are joined in the given nucleotide neighborhood type. The resulting graph is isomorphic to a 6D hypercube [13,26]. This high-dimensional cubic graph of the 64 triplets is a natural extension of the nucleotides arranged in a square. A codon graph is a graph in which the vertices represent codons and are joined according to a nucleotide neighborhood type. Codon graphs can be constructed for any subset of the 64 possible triplets. The RNY code has been modeled as a 4D hypercube [26,29,30]. Two genetic codes from which the primeval RNA code [25] could have originated the SGC have been derived [26]. Given the RNY code, the necessary transformations that are needed to obtain the SGC are simple algebraic operations: rotations (for the Extended RNA code type I) and translations (for the Extended RNA code type II) in the vector space $GF(4)$ in 3 dimensions [26].

The Extended RNA code type I consists of RNY, NYR and YRN codons. The extended RNA code type II comprises all codons of the type RNY, YNY and BNR [26]. Then, by performing frame-reading misreadings (Extended code I), 48 codons that specify 17 amino acids and the three stop codons are obtained. If transversions in the 1st or 3rd nucleotide bases of the RNY pattern are permitted, then there are also 48 codons that encode for 18 amino acids without stop codons, then there are also 48 codons that encode for 18 amino acids without stop codons (Extended code II). The codons in each of the subsets of both Extended RNA codes were represented by 4D symmetrical hypercubes [26], whose union comprised precisely the already-known 6D hypercube of the SGC of 64 triplets [31]. Evolutionary analysis of SGC based upon 3D algebraic models, dubbed Genetic Hotels, leads more clearly to the same conclusions [32]. The composition of both evolutionary paths yields to the complete set of 64 codons of the SGC. Mitochondrial codes present variations principally in the codons for the stop signals and uncoded amino acids. The mitochondrial genetic codes of yeast, invertebrates, and vertebrates are shown in Table 1. They were downloaded from: <https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?chapter=teencodes#SG24> (accessed on July 31, 2018). Note that in the mitochondrial genetic codes (Table 1) every amino acid has a set of coding triplets with an even number of elements. Note that Ile is tricodonic in SGC but it is dicodonic in all mitochondrial codes; Leu is tetracodonic in all codes except in yeast's mitochondria, which is dicodonic; Trp is uncoded only in SGC whilst it is tricodonic in all mitochondrial codes. Met is uncoded in SGC, but it is dicodonic in all mitochondrial codes. Ser is hexacodonic in SGC, vertebrate, and yeast mitochondria, but octacodonic in invertebrate mitochondria. The stop codons are tricodonic in SGC, tetracodonic in vertebrate mitochondria, and dicodonic in invertebrate and yeast mitochondria.

Table 1. SGC and mitochondrial codes.

Amino Acid	Standard Genetic Code		Vertebrate Mitochondrial Code		Invertebrate Mitochondrial Code		Yeast Mitochondrial Code	
Ala	GCA	GCC	GCA	GCC	GCA	GCC	GCA	GCC
	GCG	GCU	GCG	GCU	GCG	GCU	GCG	GCU
Arg	CGA	CGC	CGA	CGC	CGA	CGC	CGA	CGC
	CGG	CGU	CGG	CGU	CGG	CGU	CGG	CGU
	AGA	AGG					AGA	AGG
Asn	AAC	AAU	AAC	AAU	AAC	AAU	AAC	AAU
Asp	GAC	GAU	GAC	GAU	GAC	GAU	GAC	GAU
Cys	UGC	UGU	UGC	UGU	UGC	UGU	UGC	UGU
Gln	CAA	CAG	CAA	CAG	CAA	CAG	CAA	CAG
Glu	GAA	GAG	GAA	GAG	GAA	GAG	GAA	GAG
Gly	GGA	GGC	GGA	GGC	GGA	GGC	GGA	GGC
	GGG	GGU	GGG	GGU	GGG	GGU	GGG	GGU
His	CAC	CAU	CAC	CAU	CAC	CAU	CAC	CAU
Ile	AUA	AUC	AUC	AUU	AUC	AUU	AUC	AUU
	AUU							
Leu	UUA	UUG	UUA	UUG	UUA	UUG	UUA	UUG
	CUA	CUC	CUA	CUC	CUA	CUC		
	CUG	CUU	CUG	CUU	CUG	CUU		
Lys	AAA	AAG	AAA	AAG	AAA	AAG	AAA	AAG
Met	AUG		AUG	AUA	AUG	AUA	AUG	AUA
Phe	UUC	UUU	UUC	UUU	UUC	UUU	UUC	UUU
Pro	CCA	CCC	CCA	CCC	CCA	CCC	CCA	CCC
	CCG	CCU	CCG	CCU	CCG	CCU	CCG	CCU
Ser	UCA	UCC	UCA	UCC	UCA	UCC	UCA	UCC
	UCG	UCU	UCG	UCU	UCG	UCU	UCG	UCU
	AGC	AGU	AGC	AGU	AGC	AGU	AGC	AGU
Stop	UAA	UAG	UAA	UAG	UAA	UAG	UAA	UAG
	UGA		AGA	AGG				
Thr	ACA	ACC	ACA	ACC	ACA	ACC	ACA	ACC
	ACG	ACU	ACG	ACU	ACG	ACU	ACG	ACU
							CUA	CUC
						CUG	CUU	
Trp	UGG		UGG	UGA	UGG	UGA	UGG	UGA
Tyr	UAC	UAU	UAC	UAU	UAC	UAU	UAC	UAU
Val	GUA	GUC	GUA	GUC	GUA	GUC	GUA	GUC
	GUG	GUU	GUG	GUU	GUG	GUU	GUG	GUU

Genetic codes induce a natural partition of the codons and determine an equivalence relation. In this equivalence relation, two codons are considered equivalent if they codify for the same amino acid or stop signal. A graph and an equivalence relation can be combined to construct a quotient graph [33]. The set of vertices of the quotient graph are the equivalent classes, and two vertices of the quotient graph are joined if there are elements of the equivalence classes that are joined in the original graph. The quotient graph derived from the codon graphs and a genetic code are known as phenotypic graphs [27–30]. The phenotypic graph represents the phenotypic expression of the codon hypercube; the vertices represent the 20 amino acids and the stop signal [22].

The symmetries of the codon graphs of the RNY code, the extended codes, and the complete codes are analyzed for the four neighborhood types of the nucleotides. A description of the codon graphs is provided (Table 2). Labeling the vertices by the codons, the symmetries are analyzed by determining the automorphisms that keep invariant all the sets of equivalence classes for each of the genetic codes. The phenotypic graphs are constructed for the four genetic codes, in each of the evolutionary steps for the SGC and for the complete code for the mitochondrial codes. Loops in the phenotypic graphs are considered if there is a pair of elements in an equivalence class that are adjacent in the codon graph.

codon hypercube; the vertices represent the 20 amino acids and the stop signal [22]. The symmetries of the codon graphs of the RNY code, the extended codes, and the complete codes are analyzed for the four neighborhood types of the nucleotides. A description of the codon graphs is provided (Table 2). Labeling the vertices by the codons, the symmetries are analyzed by determining the automorphisms that keep invariant all the sets of equivalence classes for each of the genetic codes. The phenotypic graphs are constructed for the four genetic codes, in each of the evolutionary steps for the SGC and for the complete code for the mitochondrial codes. Loops in the phenotypic graphs are determined for the four genetic codes in the four neighborhood types of nucleotides. The automorphisms group of the phenotypic graphs is determined for the four genetic codes in the four neighborhood types of nucleotides.

Table 2. Graphs isomorphic to the codon graphs for the 4 nucleotide neighborhoods at different evolutionary stages. C_4 and C_6 are cyclic graphs of 4 and 6 vertices, respectively; K_4 is the complete graph with 4 vertices; Q_3 is a hypercube of 3 dimensions; P_3 is the graph path of 3 vertices; Q_4 and C_6 are cyclic graphs of 4 and 6 vertices, respectively; K_4 is the complete graph with 4 vertices; Q_3 is a hypercube of 3 dimensions; P_3 is the graph path of 3 vertices.

Codons	Neighborhood 1	Neighborhood 2	Neighborhood 3	Neighborhood 4
RNY Codons	C_4	Neighborhood 2	Neighborhood 3	Neighborhood 4
Extended code 1	Supplementary Materials I	$C_6 \square Q_2$	$C_6 \square Q_2$	Supplementary Materials I
Extended code 2	Supplementary Materials I	$Q_4 \square P_3$	$Q_4 \square P_3$	Supplementary Materials I
Complete Code	Q_6	Q_6	Q_6	$K_4 \square K_4 \square K_4$

3.3. Results

The codon graphs constructed for the RNY for the four nucleotide neighborhood types result in three different graphs (Figure 2).

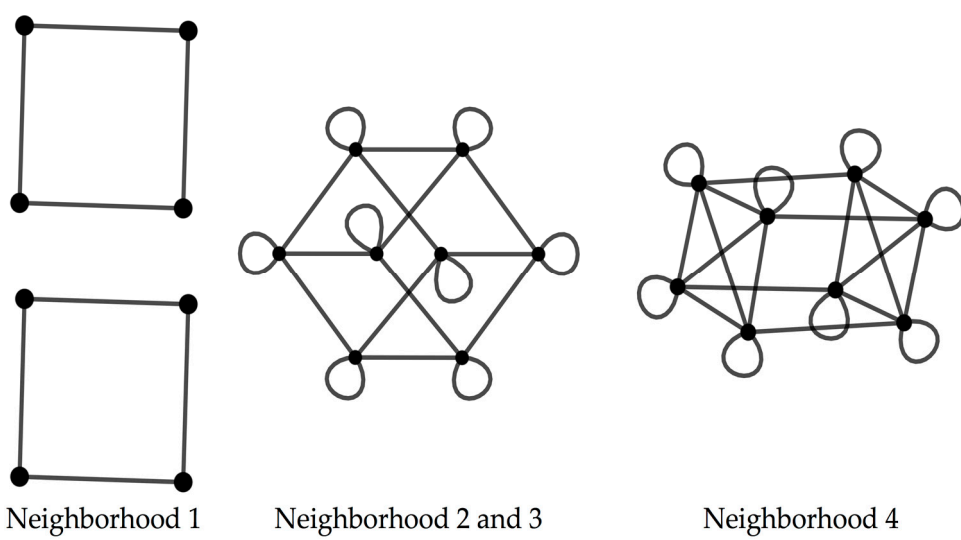


Figure 2. Codon graphs for the 4 types of neighborhoods for the RNY code.

When considering the codon graph with the vertices labeled, the neighborhood type 1 graph is composed of joints squares; the neighborhood type 2 and 3 graph is isomorphic to a 4D hypercube; the neighborhood type 4 produces a 4D hypercube with 4D unmarked diagonals, where this graph is isomorphic to the graph resulting from the Cartesian product from the Cartesian of the graphs K_4 and C_4 where K_4 is the complete graph where vertices are completely graphed for all vertices (Table 2).

For the Extended code 1, the codon graph resulting from the neighborhood type 2 and type 3 is isomorphic to the Cartesian product of the graph C_6 and Q_2 where Q_n is the hypercube of dimension n (Table 2). The matrices for the codon graphs of the Extended code 1 based on the neighborhood type 2 and type 3 are provided in Supplementary Materials I. For the codon graph from the nucleotide neighborhoods type 1 and type 4 are provided in Supplementary I.

With the vertices of the codon graphs labeled with the corresponding set of codons, these labeled codon graphs are analyzed for the four genetic codes. As these genetic codes are different, the automorphisms group that keeps invariant all the equivalent classes for each genetic code is

also different. The RNY code is the same in the SGC and the three mitochondrial codes analyzed. The corresponding automorphisms group for the four nucleotide neighborhood types are $\mathbb{Z}_2 \times \mathbb{Z}_2$ for the neighborhood type 1, and \mathbb{Z}_2 for the rest of the nucleotide neighborhood types (Table 3).

Table 3. Automorphism groups for the SGC and mitochondrial codes for the 4 types of nucleotide neighborhoods that preserve the codon sets of all the amino acids. \mathbb{Z}_2 is the binary field of 2 elements and $\mathbb{Z}_2 \times \mathbb{Z}_2 = \{00,01,10,11\}$.

Genetic Code	Codons	Neighborhood 1	Neighborhood 2	Neighborhood 3	Neighborhood 4
Standard Genetic Code	RNY code	$\mathbb{Z}_2 \times \mathbb{Z}_2$	\mathbb{Z}_2	\mathbb{Z}_2	\mathbb{Z}_2
	Extended code 1	\mathbb{Z}_2	E	E	\mathbb{Z}_2
	Extended code 2	$\mathbb{Z}_2 \times \mathbb{Z}_2$	E	E	\mathbb{Z}_2
	Complete Code	\mathbb{Z}_2	E	E	\mathbb{Z}_2
Mitochondrial Codes	Complete Code	$\mathbb{Z}_2 \times \mathbb{Z}_2$	\mathbb{Z}_2	\mathbb{Z}_2	$\mathbb{Z}_2 \times \mathbb{Z}_2$

For the Extended codes in the SGC, the nucleotide neighborhoods type 2 and type 3 possess no symmetries, as the automorphisms group is the trivial one. For the neighborhood type 1, the automorphisms groups for the Extended codes 1 and 2 are \mathbb{Z}_2 and $\mathbb{Z}_2 \times \mathbb{Z}_2$, respectively. Considering the nucleotide neighborhood type 4, the automorphisms groups for both extended codes are \mathbb{Z}_2 (Table 3). The three mitochondrial codes exhibit the same automorphisms groups in the codon graphs for the Extended codes (Table 3). For the complete code, the codon graph for the SGC only possesses a symmetry given by the group \mathbb{Z}_2 in the nucleotide neighborhoods type 1 and type 4. In the three mitochondrial codes, the codon graphs for the complete code present as automorphisms group, the group $\mathbb{Z}_2 \times \mathbb{Z}_2$ for the neighborhoods type 1 and type 4; the group \mathbb{Z}_2 is the symmetry group for the neighborhoods type 2 and type 3 (Table 3). The codon graphs for the three mitochondrial codes not only share the same amino-acid-preserving symmetries, but the elements of these groups are the same. This result shows that the codons that the three mitochondrial codes have in common, which are different from the SGC, are the source of symmetry. Specifically, the swap of the codon AUA from Ile to Met increases the symmetries of the mitochondrial codes. This codon is neighbor to the Met codon AUG in the nucleotide neighborhood types 2, 3, and 4. The codons AUA and AUG are present in both Extended codes, hence, the codon graphs for mitochondrial codes are more symmetric than the SGC. A detailed description of the automorphisms groups in permutation representation is provided in Supplementary II.

The phenotypic graphs were constructed for all the nucleotide neighborhood types, at the four evolutionary stages and for the four genetic codes analyzed. The phenotypic graphs for the RNY code present nontrivial automorphisms groups. For the nucleotide neighborhood type 1, the automorphisms group is given by $D_4 \times D_4 \times S_2$ where D_n is the dihedral group of a regular n-gon and S_n is defined as the symmetric group of n elements; for the rest of the nucleotide neighborhood types, the automorphisms group is the octahedral group O_h .

For the rest of the evolutionary stages and genetic codes, only the phenotypic graph of the complete code for the invertebrate mitochondrial code, based on the nucleotide neighborhood type 3, has as automorphisms group, the group \mathbb{Z}_2 , whereas the rest of the phenotypic graphs hold no symmetries. The reflection on the phenotypic graph for the complete invertebrate mitochondrial code on the nucleotide neighborhood type 3 is given by the permutation that interchanges the amino acids of Ile and Met. Note that the codons of these two amino acids generate the symmetries of the codon graphs for the mitochondrial codes. Phenotypic graphs for the SGC for the four nucleotide neighborhood types are shown in Figure 3

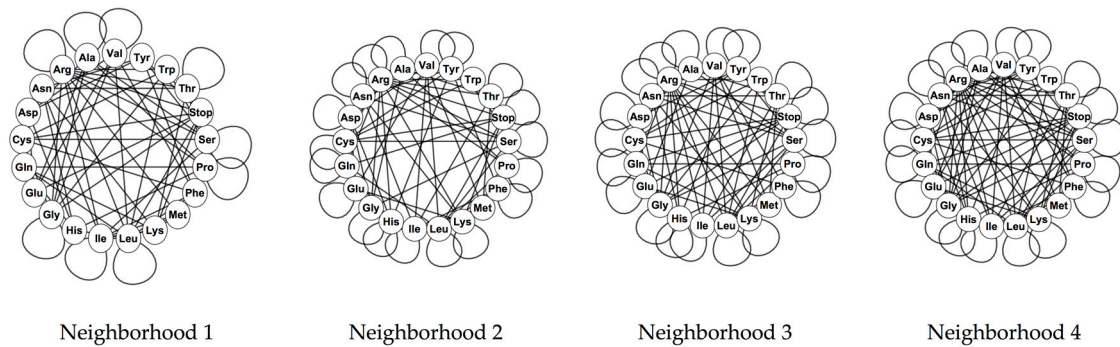


Figure 3. Phenotypic graphs of the SCC for the 4 types of nucleotide neighborhoods.

4. Discussion

In this work, we analyzed the symmetric structure of different genetic codes with graph theory. With codon graphs and phenotypic graphs, we analyzed both sides of a genetic code: the genotype and its phenotype. Our method is a novel approach to analyze any genetic code, even synthetic ones. The codon graphs allow us to analyze the structure and evolution of the genetic code through its evolutionary stages. Each nucleotide neighborhood type spans a different graph in each evolutionary step of a genetic code, both for codons and amino acids. The automorphism group of the codon graphs that keep invariant the sets of codons or amino acids reflects the automorphism group of the codons to the phenotype. The degeneracy of the genetic code is given by the words kept by the symmetric structure and the priority of each amino acid type. The degeneracy assignments of codons to amino acids properly compares with the phenotypic graphs mainly their product. This system determines codon swaps that maintain the distribution of amino acids in a genetic code. These codon swaps are possible reassignments that do not change the codification of whole sets of codons for given amino acids. The reassignment with the same codon to Met in the analyzed mitochondrial codes emerges as the source of the symmetry of the mitochondrial codes. The set of codons of vertebrate mitochondria are different from the standard codons in yeast mitochondria. Yet these differences do not explain the increase of symmetry in mitochondrial codes. We remark that despite differences among mitochondrial codes, they display the same type of symmetry. Definitely, mitochondrial codes are different among themselves and different from the standard codes, at least the symmetry observed in the SCC. Even more, they show a more symmetrical structure than the SCC and at the same time they conserve the symmetrical structure of the SCC. Even more, we proved that mitochondrial codes are more symmetrical than the SCC. Then, the Four Klein group can be found in all codes and interestingly, we also found that the Z_{12} group in the mitochondrial codes analyzed. Understanding the differences among them, we pointed out that the origin of the increase in symmetry is due to changes in the mitochondrial amino acids but not in the standard codons or in the octacodon amino acids. Our work does not allow us to discuss if the mitochondrial codes are the result of evolutionary progress or because of retrogression. What we can safely say is that changes in the mitochondrial code are restricted to certain codons and not all changes seem to be allowed. Evolving early tend to freeze the structure in the form of the standard code and having similar levels of abundance. Departures involve only a few codons, so that the structure of the code has remained almost frozen at least since the time of the UGA of the standard code (call us) life forms. These changes were adaptations that kept only a few sequences fixed to have a universal code and facilitated the diversification of living organisms. This universality of the genetic code and the maintenance of these changes were a requirement of the early sequences. The life forms that probably obeyed the Extended RNA code types and they were present in an intermediate between the RNA World and the LUCA. They pertained to the Ribonucleoprotein World. The reformer systems that are not RNA under typical state and they may show universal properties of scale invariance [35], the RNA World and LUCA. They pertained to the Ribonucleoprotein World. Therefore, genomes are systems that are constantly under a critical state and they may show universal properties of scale invariance [35].

The 6D hypercube has been used to analyze different biological properties of the SGC. Woese's [36] polar requirement property broadly distinguishes the amino acids into four categories. The polar requirement is a physico-chemical property of the amino acids and is directly associated to the organization of the SGC [37]. Polar requirement is related to the division of amino acids in a polar–nonpolar interface [38]. The relation between the assignments in the SGC and the polar requirements is reflected in the symmetrical pattern that arises when the polar requirement categories are used to color the codon graphs of the SGC [13]. Genetic codes are implemented via tRNA molecules and their anticodons. These molecules bind the codons in mRNA to their corresponding anticodons, then link the appropriate amino acids as determined by the mRNA. There are 20 different tRNAs, one for each amino acid. A tRNA is charged with its corresponding amino acid with the action of the aaRSs. The aaRSs are divided into two families, class I and class II, according to the groove of tRNA with which they interact, minor groove or major groove. The Rodin–Ohno model of the genetic codes divides the codon table into two categories by which class of aaRSs is responsible to charge the amino acid associated with each codon [39,40]. The division of the SGC table by the two classes of aaRSs was argued as “almost symmetrical” [40], although, this symmetrical partition of the SGC was shown with the codon graphs of the SGC [13].

Given the set of 64 codons that codify for 20 canonical amino acids and a stop signal, there are $21^{64} \approx 4 \times 10^{84}$ possible genetic codes. This calculation does not assume the evolution and degeneracy of the SGC. Coupling codon graphs of the SGC with different biological properties have allowed the analysis of several biological properties that uniquely determine the current SGC [29].

The robustness and optimality of the SGC have been widely analyzed [30,41–43] and found suboptimal according to its error correction properties. Phenotypic graphs of random codes that maintain specific properties of the SGC have been analyzed for their connectivity properties. It was shown that despite the current SGC being suboptimal (regarding error tolerance, for example), it is optimal if its evolutionary history is considered [30]. For the SGC to reach its optimal state of error tolerance, it would require codon swaps that are evolutionarily incompatible as these paths fix the SGC in each stage.

Other nucleotide models represent them by using a bijection from the nucleotides to the elements of the Galois field of four elements $GF(4)$ [17,27,32]. With this bijection, an algebraic structure is given to the nucleotides. Representing the field $GF(4)$ with the integer numbers from one to four, it is possible to represent the nucleotides in the real line \mathbb{R} and the codons in the space \mathbb{R}^3 . There are 24 possible assignments of the elements of the $GF(4)$ to the set $\{1,2,3,4\}$ [17]. These representations of the genetic code have been widely studied for their biological and mathematical properties [17,27,32]. Phenotypic graphs of these 3D representations have been constructed to analyze the SGC and compare it with the human tRNA code and the standard tRNA code for its centrality measures, and the role of the stop codons and different degeneracy patterns have been described [27]. Representations of the primeval RNY code have been constructed based on the bijection to the $GF(4)$ and their phenotypic graphs have been derived and analyzed for their symmetries based on polar requirement [28]. Recall that phenotypic graphs can be constructed from any graph representation of the 64 codons, or any subset of it. The graph representation of the nucleotides in a square generalizes the bijection to the $GF(4)$ and allows using group actions that represent the biological mutations, transitions and transversions, to represent the symmetries of the genetic code [13].

The two evolutionary paths arise from transformations of the primeval RNY code based on mistranslations on the early translation mechanisms and mutations on this small set of codons [26]. Geometrically, these extended codes arise from symmetry breakings and translations of an RNY four-dimensional hypercube [26,32]. The composition of both evolutionary paths completes the set of 64 codons of a genetic code.

The codon graphs constitute a useful approach to analyze the evolvability of the genetic code. All in all, the codon graphs and their derived phenotypic graphs constitute a mathematical

framework to theoretically analyze the SGC, the mitochondrial code, or any noncanonical code, including custom-designed codes.

Supplementary Materials: The following are available online at <http://www.mdpi.com/2073-8994/10/9/388/s1>.

Author Contributions: G.S.Z. and M.V.J. conceived the whole work, contributed with ideas, and wrote the manuscript; G.S.Z. performed the analyses; M.V.J. coordinated the research.

Funding: M.V.J. was funded by Dirección General de Asuntos del Personal Académico (DGAPA), Universidad Nacional Autónoma de México, UNAM (PAPIIT-IN224015); G.S.Z. is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and received a doctoral fellowship from CONACYT (number: 737920).

Acknowledgments: We thank the reviewers' comments. We thank Juan R. Bobadilla for technical computer support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Watson, J.D.; Crick, F.H.C. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* **1953**, *171*, 737–738. [[CrossRef](#)] [[PubMed](#)]
2. Matthaei, J.H.; Nirenberg, M.W. Characteristics and stabilization of DNAase-sensitive protein synthesis in *E. coli* extracts. *Proc. Natl. Acad. Sci. USA* **1961**, *47*, 1580–1588. [[CrossRef](#)] [[PubMed](#)]
3. Nirenberg, M.W.; Matthaei, J.H. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. USA* **1961**, *47*, 1588–1602. [[CrossRef](#)] [[PubMed](#)]
4. Crick, F.H.C. The origin of the genetic code. *J. Mol. Biol.* **1968**, *38*, 367–379. [[CrossRef](#)]
5. Eigen, M.; Lindemann, B.; Tietze, M.; Winkler-Oswatitsch, R.; Dress, A.; von Haeseler, A. How old is the genetic code? Statistical geometry of tRNA provides an answer. *Science* **1989**, *244*, 673–679. [[CrossRef](#)] [[PubMed](#)]
6. Nicholas, H.B.; McClain, W.H. Searching tRNA sequences for relatedness to aminoacyl-tRNA synthetase families. *J. Mol. Evol.* **1995**, *40*, 482–486. [[CrossRef](#)] [[PubMed](#)]
7. Delarue, M. An asymmetric underlying rule in the assignment of codons: possible clue to a quick early evolution of the genetic code via successive binary choices. *RNA* **2007**, *13*, 161–169. [[CrossRef](#)] [[PubMed](#)]
8. Ribas de Pouplana, L.; Schimmel, P. Aminoacyl-tRNA synthetases: Potential markers of genetic code development. *Trends Biochem. Sci.* **2001**, *26*, 591–596. [[CrossRef](#)]
9. Rodin, S.N.; Rodin, A.S. On the origin of the genetic code: signatures of its primordial complementarity in tRNAs and aminoacyl-tRNA synthetases. *Heredity (Edinb.)* **2008**, *100*, 341–355. [[CrossRef](#)] [[PubMed](#)]
10. Eriani, G.; Delarue, M.; Poch, O.; Gangloff, J.; Moras, D. Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. *Nature* **1990**, *347*, 203–206. [[CrossRef](#)] [[PubMed](#)]
11. Hornos, J.E.M.; Hornos, Y.M.M. Algebraic model for the evolution of the genetic code. *Phys. Rev. Lett.* **1993**, *71*, 4401–4404. [[CrossRef](#)] [[PubMed](#)]
12. Sánchez, R.; Morgado, E.R.; Grau, R. A genetic code Boolean structure. I. The meaning of Boolean deductions. *Bull. Math. Biol.* **2005**, *67*, 1–14. [[CrossRef](#)] [[PubMed](#)]
13. José, M.V.; Zamudio, G.S.; Morgado, E.R. A unified model of the standard genetic code. *R. Soc. Open Sci.* **2017**, *4*, 160908. [[CrossRef](#)] [[PubMed](#)]
14. Crick, F.H.C. On Protein Synthesis. *Symp. Soc. Exp. Biol.* **1958**, *12*, 138–166. [[PubMed](#)]
15. Crick, F.H.C.; Brenner, S.; Klug, A.; Piecznik, G. A speculation on the origin of protein synthesis. *Orig. Life* **1976**, *7*, 389–397. [[CrossRef](#)] [[PubMed](#)]
16. Crick, F.H.C. Codon-anticodon pairing: The wobble hypothesis. *J. Mol. Biol.* **1966**, *19*, 548–555. [[CrossRef](#)]
17. José, M.V.; Morgado, E.R.; Sanchez, R.; Govezensky, T. The 24 possible algebraic representations of the standard genetic code in six or in three dimensions. *Adv. Stud. Biol.* **2012**, *4*, 119–152.
18. Voet, D.; Voet, J.G.; Pratt, C.W. *Fundamentals of Biochemistry: Life at the Molecular Level*; Wiley: Hoboken, NJ, USA, 2012; ISBN 0470547847.
19. Logan, D.C. The mitochondrial compartment. *J. Exp. Bot.* **2007**, *58*, 1225–1243. [[CrossRef](#)] [[PubMed](#)]
20. Suen, D.F.; Norris, K.L.; Youle, R.J. Mitochondrial dynamics and apoptosis. *Genes Dev.* **2008**, *22*, 1577–1590. [[CrossRef](#)] [[PubMed](#)]

21. Tait, S.W.G.; Green, D.R. Mitochondria and cell signalling. *J. Cell Sci.* **2012**, *125*, 807–815. [[CrossRef](#)] [[PubMed](#)]
22. Osawa, S.; Jukes, T.H.; Watanabe, K.; Muto, A. Recent-Evidence for Evolution of the Genetic-Code. *Microbiol. Rev.* **1992**, *56*, 229–264. [[PubMed](#)]
23. Ambrogelly, A.; Palioura, S.; Söll, D. Natural expansion of the genetic code. *Nat. Chem. Biol.* **2007**, *3*, 29–35. [[CrossRef](#)] [[PubMed](#)]
24. Beardon, A.F. *Algebra and Geometry*; Cambridge University Press: New York, NY, USA, 2005.
25. Eigen, M.; Schuster, P. The Hypercycle-A principle of natural self-organization Part B: The abstract hypercycle. *Naturwissenschaften* **1978**, *65*, 7–41. [[CrossRef](#)]
26. José, M.V.; Morgado, E.R.; Govezensky, T. An extended RNA code and its relationship to the standard genetic code: an algebraic and geometrical approach. *Bull. Math. Biol.* **2007**, *69*, 215–243. [[CrossRef](#)] [[PubMed](#)]
27. José, M.V.; Morgado, E.R.; Guimarães, R.C.; Zamudio, G.S.; de Fariás, S.T.; Bobadilla, J.R.; Sosa, D. Three-Dimensional Algebraic Models of the tRNA Code and 12 Graphs for Representing the Amino Acids. *Life* **2014**, *4*, 341–373. [[CrossRef](#)] [[PubMed](#)]
28. José, M.V.; Zamudio, G.S.; Palacios-Pérez, M.; Bobadilla, J.R.; de Fariás, S.T. Symmetrical and Thermodynamic Properties of Phenotypic Graphs of Amino Acids Encoded by the Primeval RNY Code. *Orig. Life Evol. Biosph.* **2015**, *45*, 77–83. [[CrossRef](#)] [[PubMed](#)]
29. Zamudio, G.S.; José, M.V. On the Uniqueness of the Standard Genetic Code. *Life* **2017**, *7*, 7. [[CrossRef](#)] [[PubMed](#)]
30. Zamudio, G.S.; José, M.V. Phenotypic Graphs and Evolution Unfold the Standard Genetic Code as the Optimal. *Orig. Life Evol. Biosph.* **2018**, *48*, 83–91. [[CrossRef](#)] [[PubMed](#)]
31. Jiménez-Montaño, M.A.; de la Mora-Basáñez, C.R.; Pöschel, T. The hypercube structure of the genetic code explains conservative and non-conservative aminoacid substitutions in vivo and in vitro. *Biosystems* **1996**, *39*, 117–125. [[CrossRef](#)]
32. José, M.V.; Morgado, E.R.; Govezensky, T. Genetic hotels for the standard genetic code: evolutionary analysis based upon novel three-dimensional algebraic models. *Bull. Math. Biol.* **2011**, *73*, 1443–1476. [[CrossRef](#)] [[PubMed](#)]
33. Auer, B.; Bisseling, R. *Graph Partitioning and Graph Clustering*; Bader, D., Meyerhenke, H., Sanders, P., Wagner, D., Eds.; Contemporary Mathematics; American Mathematical Society: Providence, RI, USA, 2013.
34. Imrich, W.; Klavžar, S.; Rall, D.F. *Topics in Graph Theory. Graphs and Their Cartesian Product*; AK Peters/CRC Press: New York, NY, USA, 2008.
35. José, M.V.; Govezensky, T.; García, J.A.; Bobadilla, J.R. On the evolution of the standard genetic code: Vestiges of critical scale invariance from the RNA world in current prokaryote genomes. *PLoS ONE* **2009**, *4*. [[CrossRef](#)] [[PubMed](#)]
36. Woese, C.R.; Dugre, D.H.; Saxinger, W.C.; Dugre, S.A. The molecular basis for the genetic code. *Proc. Natl. Acad. Sci. USA* **1966**, *55*, 966–974. [[CrossRef](#)] [[PubMed](#)]
37. Alff-Steinberger, C. The genetic code and error transmission. *Proc. Natl. Acad. Sci. USA* **1969**, *64*, 584–591. [[CrossRef](#)] [[PubMed](#)]
38. Mathew, D.C.; Luthey-Schulten, Z. On the physical basis of the amino acid polar requirement. *J. Mol. Evol.* **2008**, *66*, 519–528. [[CrossRef](#)] [[PubMed](#)]
39. Carter, C.W.; Li, L.; Weinreb, V.; Collier, M.; Gonzalez-Rivera, K.; Jimenez-Rodriguez, M.; Erdogan, O.; Kuhlman, B.; Ambroggio, X.; Williams, T.; et al. The Rodin-Ohno hypothesis that two enzyme superfamilies descended from one ancestral gene: an unlikely scenario for the origins of translation that will not be dismissed. *Biol. Direct* **2014**, *9*, 11. [[CrossRef](#)] [[PubMed](#)]
40. Rodin, A.S.; Szathmáry, E.; Rodin, S.N. On origin of genetic code and tRNA before translation. *Biol. Direct* **2011**, *6*, 14. [[CrossRef](#)] [[PubMed](#)]
41. Novozhilov, A.S.; Wolf, Y.I.; Koonin, E.V. Evolution of the genetic code: Partial optimization of a random code for robustness to translation error in a rugged fitness landscape. *Biol. Direct* **2007**, *2*, 24. [[CrossRef](#)] [[PubMed](#)]

42. Haig, D.; Hurst, L.D. A Quantitative Measure of Error Minimization in the Genetic-Code. *J. Mol. Evol.* **1991**, *33*, 412–417. [[CrossRef](#)] [[PubMed](#)]
43. Freeland, S.J.; Knight, R.D.; Landweber, L.F.; Hurst, L.D. Early Fixation of an Optimal Genetic Code. *Mol. Biol. Evol.* **2000**, *17*, 511–518. [[CrossRef](#)] [[PubMed](#)]



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Copyright of Symmetry (20738994) is the property of MDPI Publishing and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.

Apéndice 5

Identity Elements of tRNA as Derived from Information Analysis

Gabriel S. Zamudio¹ · Marco V. José¹

Received: 3 April 2017 / Accepted: 9 June 2017 /

Published online: 28 June 2017

© Springer Science+Business Media B.V. 2017

Abstract The decipherment of the tRNA's operational code, known as the identity problem, requires the location of the sites in the tRNA structure that are involved in their correct recognition by the corresponding aminoacyl-tRNA synthetase. In this work, we determine the identity elements of each tRNA isoacceptor by means of the variation of information measure from information theory. We show that all isoacceptors exhibit sites associated with some bases of the anticodon. These sites form clusters that are scattered along the tRNA structure. The clusters determine the identity elements of each tRNA. We derive a catalogue of clustered sites for each tRNA that expands previously reported elements.

Keywords Identity elements · Operational code · Anticodon code · tRNA evolution · Information theory

Introduction

The correct implementation of the genetic code comprehends an extensive and complex set of biological interactions inside the cell. At the core of the system lies the transfer RNA (tRNA), a key molecule driving the translation process. To preserve this intricate process, the tRNA is equipped with two codes along its structure. The anticodon code that reads the codons of a messenger RNA (mRNA), and a “second” “operational code (De Duve 1988) which is commonly associated to the acceptor stem of the tRNA (Hou and Schimmel 1988). Since the dawn of the mini-helix structure of tRNA (Tamura 2015) 3.5 billion years ago, both codes have coevolved and they are encrypted in its structure. tRNA molecules share a common

Electronic supplementary material The online version of this article (doi:10.1007/s11084-017-9541-6) contains supplementary material, which is available to authorized users.

✉ Marco V. José
marcojose@biomedicas.unam.mx

¹ Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, CDMX, C.P. 04510 Ciudad de México, Mexico

structure that is recognized by other molecules of the biosynthetic pathways (Arnez and Moras 1997). In the translation process, the tRNA is the only participant which recognizes the codons of a mRNA directly (Cusack 1997). A mature tRNA has a canonical length of 76 bases (Altman 1993), although some isoacceptors (tRNAs with different anticodons for the same amino acid) differ in length, mainly at the variable loop. The anticodon triplet is located at bases 34, 35, and 36. The terminal CCA motif at 3' end, which is added post-transcriptionally (Hou 2010), is the place at which amino acids are attached through an esterification reaction with its cognate aminoacyl-tRNA synthetase (aaRS) that has been previously charged with an amino acid (Arnez and Moras 1997). The “second” genetic code, also called “operational code” (De Duve 1988; Rodin and Ohno 1997), directs the correct identification of tRNA isoacceptors with its respective aaRSs by stereochemical means. The problem of deciphering the operational code is known as “the identity problem”, since the molecules of aaRS must identify the correct tRNA from a pool of similar molecules, without necessarily interacting with the anticodons.

Two protein subfamilies compose the set of aaRSs, Class I, and Class II, with ten proteins each (Eriani et al. 1990). The 20 different aaRSs account for the 20 canonical amino acids, and so, the operational code is nondegenerate (Eriani et al. 1990). In contrast the anticodon code consists of 48 anticodons, since triplets starting with adenine are absent of this code (Guimarães et al. 2008; José et al. 2014). The two classes of aaRSs recognize the acceptor helix of the tRNAs by different approaches: Class I recognizes the minor groove and charges the amino acid at the 2'OH group of the terminal adenosine, while Class II access from the major groove and charges the amino acid in the 3'OH group (Eriani et al. 1990). It has been shown that the specificities between tRNAs and aaRSs coevolved during the formation of the genetic code and they were driven by the hydrophathy of anticodons (Farias et al. 2014a).

Information theory has been used to analyze genetic sequences (Adami 2004, 2012). It has also been used to predict the secondary structure of RNA (Durbin et al. 1998). In this work, we use the measure of variation of information, from information theory, to determine the specific sites in the tRNA structure that are highly related to the anticodon. This measure uses the variation in the gene sequences of a tRNA isoacceptor to locate the sites that contribute to the degeneracy of the isoacceptor's anticodon code. These identity elements determine the recognition process of tRNAs by their respective aaRSs in the translation process.

Data Sources

The database (Abe et al. 2014) contains curated tRNA genes from the three kingdoms of life. We selected those sequences whose lengths matched the canonical length of 76 nucleotides discarding the variable loop. Redundant sequences were also omitted in order to avoid duplicated sequences that could alter the statistical results. Overall, we analyzed a total of 13,093 gene sequences including all isoacceptors.

Methods

Information theory is devoted to the quantification, transmission and storage of information. In particular, given two messages, it is possible to determine the information shared by them, and consequently compare them. The variation of information is a measure that determines the

information distance between two messages. This measure is used as a clustering algorithm for data and for comparing different clusters, of the same data (Meila 2003, 2007). The variation of information is a measure that gives a distance between two messages X and Y . It is given by the equation $VI(X, Y) = H(X) + H(Y) - 2I(X, Y)$ where $H(X)$ is the Shannon's entropy, and the term $I(X, Y)$ accounts for the mutual information shared between X and Y (Meila 2003). The random variables X and Y , describe the distribution of characters or symbols in any given message. The results are given in bit units. This measure captures the information needed to describe one variable from previously knowing the other. As the mutual information function is a factor of the variation of information, the pairs of sites close to each other have a general dependence, not necessarily a linear dependence that can be obtained with the correlation function (Li 1990). Dividing the tRNA genes by isoacceptors, and using the variation of information per site, the distribution of nucleotides in a single site of the tRNA was considered as a random variable. Then, it was possible to compute the variation of information distance between any two sites of the same isoacceptor, for each of the 20 amino acids. For the calculations, the terminal CCA was removed from the sequences, as it is a constant motif in all of them. For small sample sizes a correction is applied to the entropy and mutual information function to account for the bias. The approximated error is based on the number of states and the sample size (Li 1990). For the variation of information, the error is approximated by $\overline{VI(X, Y)} - VI(X, Y) \approx \frac{K^2}{N}$, where $\overline{VI(X, Y)}$ stands for the true variation of information, while $VI(X, Y)$ is the calculated variation of information, the number of states is denoted by K and N is the sample size. This error is applied uniformly to all calculated distances. As the error is applied for each isoacceptor, regardless of the sample size, it can be neglected. If the error is considered, the methodology would require obtaining the minimum distances and the same results would be attained, since this distance would coincide with the estimated error.

Results

If two sites x_1 and x_2 are at distance zero, it means that those sites are clustered and so, the state or occurrence of a base in the site x_1 is completely predictable by the state of x_2 , and vice versa. Hence, the sites involved with the correct aminoacylation of tRNAs, i.e., the identity elements, would be those whose distance with any base with the anticodon is zero. These sites form clusters that are involved in the proper recognition of the corresponding aaRSs. The nucleotide bases present in the sites of a cluster are coordinated. This means that each nucleotide present in a site is derived by the presence of a nucleotide found in another site of the cluster. Table 1 enlists the sets of sites that are fully clustered forming a unit dividing the isoacceptors by its aaRSs class; anticodon positions are marked in red. Notice that some isoacceptors contain multiple sets of predictable sites. Also notice that all the sites in a single set are at distance zero of each other. The presence of multiple clusters in an individual isoacceptor is apparent. Some form Watson-Crick pairs (marked in green) which are relevant to maintain the stability of the secondary and tertiary structure of the tRNA molecules. Other clusters are used to avoid misidentification and mischarging of the tRNA (Giegé et al. 1998). As an example, the clusters for tRNA^{Gln} are colored on the secondary structure (Fig. 1). The sites that belong to the same cluster set are colored with the same color. Notice that the tRNA^{Gln} possesses three clusters: i) The cluster with the 35, 36 anticodon bases associated with five bases at the TΨC-loop; ii) a

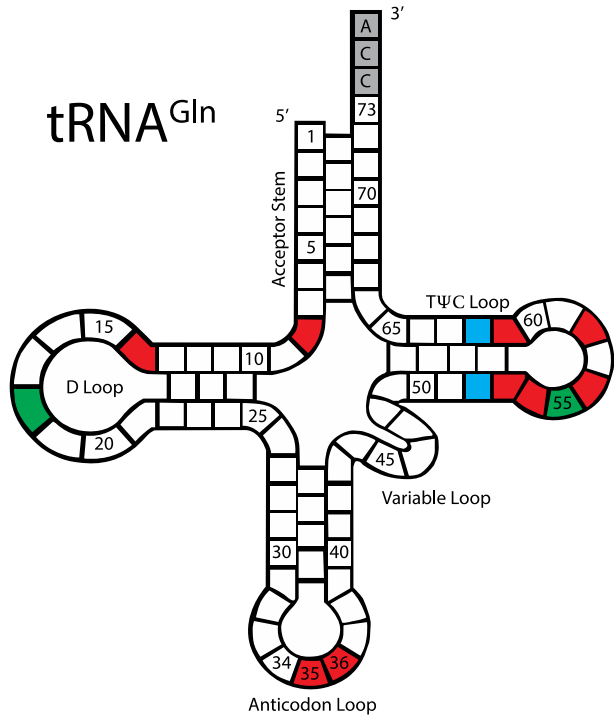
Table 1 Table of clusters for each tRNA. Each set conforms a cluster of positions. Anticodon bases are in red. Positions forming Watson-Crick pairs are marked in green. The sites that do not present a Watson-Crick pair are marked in black. The symbol Nstands for the sample size of each isoacceptor

tRNA	Class I	
	Unity Clusters	N
tRNA ^{Arg}	{8,14, 35 ,54,58}	1249
tRNA ^{Cys}	{1,8,14,18, 34,35,36 ,53,54,56,58,61,72}, {2,71}, {3,70}, {5,68}, {7,66}, {10,25}, {52,62}	131
tRNA ^{Gln}	{8,14, 35,36 ,53,54,56,58,61}, {18,55}, {52,62}	354
tRNA ^{Glu}	{ 35,36 ,58}, {53,61}	721
tRNA ^{Ile}	{8,10,14,18,22,25, 35,36 ,53,54,56,58,61}, {7,66}, {11,24}, {12,23}, {28,42}, {29,41}, {50,64}	32
tRNA ^{Leu}	{8,14,18, 35 ,53,54,56,58,61}, {52,62}	731
tRNA ^{Met}	{8,33, 34,35,36 ,53,54,56,58,61}	559
tRNA ^{Phe}	{8,10,12,21,23,33, 36 ,53,56}, {34,35}	812
tRNA ^{Trp}	{18,33, 35,36 ,37,53,55,58,61}	552
tRNA ^{Val}	{18, 35,36 }, {12,23}	774

tRNA	Class II	
	Unity Clusters	N
tRNA ^{Ala}	{18, 35,36 ,54,58}, {53,61}	1563
tRNA ^{Asn}	{ 35,36 ,53,58}, {12,23}, {18,56}, {52,62}, {31,39}	579
tRNA ^{Asp}	{1,17,33, 34,35,36 ,53,54,58,61,72}, {11,24}, {50,64}, {51,63}	126
tRNA ^{Gly}	{33, 35,36 ,54,58}, {50,64}, {53,61}	804
tRNA ^{His}	{18, 34,35,36 ,53,54,55,56,61}, {12,23}, {29,41}, {51,63}	584
tRNA ^{Lys}	{18,33, 35,36 }, {37,38}, {53,61}	1490
tRNA ^{Pro}	{ 35,36 ,53,54,55,58,61,73}, {1,72}, {2,71}, {3,70}, {18,33}, {29,41}	179
tRNA ^{Ser}	{1,33,53,54,55,56,60,61}, { 35,36 }, {5,68}, {29,41}, {52,62}	55
tRNA ^{Thr}	{ 35,36 ,37}, {3,70}, {53,61}	1779
tRNA ^{Tyr}	{18,21,32,33, 34,35,36 ,53,54,56,58,60,61,73}, {30,40,55}, {2,71}, {3,70}, {4,69}, {5,68}, {8,14}, {10,25}, {12,23}, {28,42}, {31,39}, {49,65}, {52,62}	19

base in the D-loop; the base 8 that links the acceptor stem and the D-loop. The second cluster is a Watson-Crick pair at the TΨC-loop. A third cluster displays bases at opposite sides of the tRNA, the base 18 at the D-loop, and the nucleotide 55 in the TΨC-loop. Altogether, these 3 clusters represent the identity elements of the operational code for tRNA^{Gln}. The isoacceptors that correspond to Class I more generally present the nucleotide 8 from the D-loop, and in some cases the base 18 is also present. Class I often present the nucleotide 8 from the D-loop clustered with the anticodon, along with base 14. In both classes, the sites 53, 54, 58, and 61 from the TΨC-loop, are all generally associated with the anticodon when some base from that side is present, marking a diffuse pattern. The tRNA with the largest number of clusters is tRNA^{Tyr} but this observation must be taken with caution due to the small sample size. We strongly recommend increasing the sample size for tRNA^{Tyr}. The wobbling base 34 is present only in Cys, Met, Phe, Asp, His, and Tyr. There is no relation between the codonicity and the location of the identity elements. The remaining figures for all isoacceptors can be found in Appendix A.

Fig. 1 Identity elements of tRNA^{Gln}. The figure portrays the tRNA secondary structure of tRNA^{Gln}. Clusters are marked with different colors



Discussion

The present work expands the current catalogue of identity elements of the 20 canonical tRNA isoacceptor groups. Every isoacceptor possess a set of sites (clusters) that includes at least one of the anticodon bases. This is in agreement with the association of the anticodon as an identity element for all tRNAs (Giegé et al. 1998). The sites related to the anticodon are present along all the structure and are different for each tRNA group. Thus, our results are in agreement with the idiosyncratic hypothesis of identity elements (Loftfield 1972) and with its distribution on the molecule (Goddard 1977). The information theoretical approach for detecting identity elements was initiated by Durbin et al. (2002) and followed by Adami (2004, 2012). The identity elements found in yeast tRNA^{Phe} by Durbin et al. (2002), using the mutual information function, are practically the same as the ones reported in the present work.

The hypothesis of the anticodon as common regulator was later rejected with in-vitro experiments on tRNA^{Ser} that showed no evidence of recognition between the anticodon and the corresponding aaRS (Sundaharadas et al. 1968). Data from yeast and *E. coli* provided clues about the presence of identity elements in the acceptor stem, the position 73 (which is the last before the CCA), the anticodon, the variable loop and the D stem (Goddard 1977). Such results showed no specific sites that could answer the current hypothesis of universal recognition sites, hence, the sites should be idiosyncratic for each isoacceptor (Loftfield 1972). Later, experiments revealed the discriminator base pair G3:U70, which was involved in the correct recognition of tRNA^{Ala} (Vargas-Rodriguez and Musier-Forsyth 2014). Further experimental work on particular species, consisted in single modifications of nucleotides along the whole tRNA in order to detect concrete positions that decrease the aminoacylation reaction, both in-

vivo and in-vitro (Giegé et al. 1998). These experiments revealed that each tRNA isoacceptor holds specific sites involved in their correct recognition. The discriminator site 73, was recognized as an identity element in conjunction with the anticodon bases. This base is present at the anticodon cluster for the tRNA^{Pro} and tRNA^{Tyr} isoacceptors. This base has been reported for Tyr (Bonfond et al. 2005). The long variable loop of tRNA^{Ser} has been reported to be determinant for its correct recognition (Wu and Gross, 1993). It has been reported the existence of positions for positive or negative recognition that participate in the correct aminoacylation or that prevent false recognition and mischarging, respectively (Giegé et al. 1998). There are also sites with different forces of recognition, being strong or weak sites. It has also been reported that in a single organism, tRNAs from different isoaccepting groups are more similar to each other than to their isoaccepting counterparts (Saks et al. 1994). They argued that this could be due to an accumulation of neutral mutations that are blind to tRNA recognition. It has also been shown that anticodon mutations could lead to changes in the isoacceptor group and they are highly tolerated. Bioinformatic analysis has shown the discriminator base pair 1:72 for tRNA^{Tyr} and tRNA^{Trp} (Mukai et al. 2017). Some clusters that contain the anticodon are accompanied by the bases 33 and 37. This has been proposed as an extended anticodon since the bases surrounding the anticodon are recognized by the corresponding aaRSs (Yarus 1982). The entropy per site of each tRNA has been calculated and it has been shown that, in average, there is no difference between the entropy profiles between the major groove and the minor groove (José et al. 2016). This result is in agreement with the sense-antisense complementary hypothesis of a common origin of tRNAs (Rodin and Rodin 2008; Carter et al. 2014). The division of genetic code, by the aaRSs class, has been shown to be an important factor in its evolutionary process (Rodin and Ohno 1977, Carter et al. 2014; Zamudio and José 2017; José et al. 2017),

Our informational approach provides a new insight for determining the identity elements of the operational code. Our results suggest further experimental work for testing the proposed sets of identity elements.

It is widely accepted the early relevance of the acceptor mini-helix in the evolutionary development of tRNA molecules (Schimmel et al. 1993; Schimmel 1995; Rodin et al. 1996). It has been proposed that the amino acid-accepting stem emerged before the anticodon loop of tRNAs, so that the first codification obeyed an operational code where amino acids were attached to their respective tRNA without the need of anticodon loop recognition (Park and Schimmel 1988; Hou and Schimmel 1988; Schimmel et al. 1993; Ribas de Pouplana et al. 1998). Indeed, the origin of the operational code is directly related to the absence of the anticodon loop in tRNAs, which enabled the first peptides to be synthesized in the absence of a genetic code (Belousoff et al. 2010). It has also been suggested that the primitive ribosome worked to synthesize peptides randomly, without the need of a code (Belousoff et al. 2010). However, Shimizu (1995) showed that small tRNAs with the portion of the anticodon loop could bind the amino acid as well. Hence, the anticodon loop was already present in primitive tRNA and should have been important in establishing specific interactions between tRNAs and their corresponding aminoacyl-tRNA synthetases.

Farias et al. (2014a, b) suggested that the coding system was assembled by co-evolution between tRNAs and aminoacyl-tRNA synthetases, being driven by changes in the second base of the anticodon of tRNA, which in turn changed the hydrophathy of the anticodon, and this pressure guided the diversification of aminoacyl-tRNA synthetases. Therefore, the recognition of the anticodon was shown to be essential for the development of the encoding system and acted as a selective pressure for the diversification of aaRSs. The two domains of the L-shaped

tRNA would have arisen independently, with the acceptor branch appearing first. In a later stage in history, the catalytic cores of synthetases emerged independently in their class I and II versions. Co-evolution of catalytic cores of synthetases and accepting hairpins led to an operational RNA code that associated specific amino acids with hairpin structures (Park and Schimmel 1988; Schimmel 1995). The anticodon domain of tRNA and the additional domains of synthetases appeared later in evolution. Anticodon domains brought the link between the RNA operational code and the correlated tRNA recognition by synthetases with the anticodon-dependent recognition by mRNA.

It has been shown (Farias et al. 2017) that the initial existence of an operational code was due to the agglutination capacity of tRNAs without the presence of a genetic code in the Peptidyl Transferase Center (PTC). This suggests that the anticodon loop initially increased the specificity between tRNAs and amino acids, and after the emergence of the proto-genes, by an exaptation process, the anticodon loops were co-opted to interact with the proto-genes, and thus the genetic code emerged and started to decode the biological information. In this manner, the emergence of the operational code and the genetic code occurred simultaneously and both systems played complementary roles in the origin and evolution of the translation system (Farias et al. 2017).

The peptides synthesis without the need of a code, indicates that in the early stages of the process the anticodon had other functions: for example, they could have been involved in the establishment of the assignments between amino acids and tRNAs. Thus, the emergence of the first genes or proto-genes must have reorganized the modes of interaction between tRNAs and PTC, which defined the sites of interaction A, P, and E in the modern ribosome. Hence, the anticodons were co-opted to stabilize the binding with the proto-gene and from this secondary interaction, the genetic correspondences between codons and amino acids were gradually established (Farias et al. 2017). The emergence of the genetic code must have occurred as an exaptation process, where anticodons would initially have the function of increasing the specificity between amino acids and tRNAs. The appearance of proto-genes would have been co-opted to establish a correlation between the information contained in the nucleic acids to proteins. The origin of the translation system is a major evolutionary transition because it enabled the establishment of the ribonucleoprotein world. The tRNAs molecules played a central role during the origin of translation, bridging RNA and (RNA + proteins) worlds. These tRNAs may offer clues towards the elucidation of the origin of the genetic code. Farias et al. (2014a, b) reconstructed the ancestral sequences of tRNAs, and when they compared concatamers of these tRNAs with the sequence of PTC from *Thermus thermophilus*, a similarity of 50.52% was found. Therefore, they suggested that the PTC arose from the junction of proto-tRNAs. It has been shown that the three-dimensional structure of the ancestral PTC, built by concatamers of ancestral sequences of tRNAs, had interactions with tRNAs anticodon loop (Farias et al. 2017).

The tRNA core hypothesis accounts for the transition from the RNA world to the Ribonucleoprotein world, where proto-tRNAs molecules possessing similar folds to those observed in modern tRNAs guided the evolutionary process of the genetic code and the translation, and enabled its fixation (Farias et al. 2014a, b). The tRNA core hypothesis (Farias et al. 2014a, b) places the tRNA at the origin of translation, i.e., tRNA molecules played a significant role in the organization of the first codified biological system, established the information storage system, and participated in coding and decoding this information. They were the protagonists of the translation system via chemical interactions with amino acids, which are inherent to the transition between the RNA world to a ribonucleoprotein world. In the tRNA core hypothesis (Farias et al. 2016), the first genes were derived from tRNA by

structural changes (tRNA-like-mRNA structure), and they enabled other tRNAs (cloverleaf tRNA canonical structure) to bind this sequence to the loop of the proto-anticodon. The tRNA molecules with the proto-tRNA (anticodon stem/loop) could interact with the amino acids (as cofactor or riboswitches), which were present in the prebiotic environment; hence, the binding between tRNA and amino acids was established. The tRNAs binding to amino acids could interact with other tRNAs in open conformation (tRNA-like mRNA structure). This interaction stabilized the complex cloverleaf tRNA canonical structure-tRNA-like mRNA structure.

Acknowledgements Gabriel S. Zamudio is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and a fellowship recipient from Consejo Nacional de Ciencia y Tecnología (CONACYT) (number: 737920). Marco V. José was financially supported by PAPIIT-IN224015, UNAM, México.

References

- Abe T, Inokuchi H, Yamada Y et al (2014) tRNADB-CE: tRNA gene database well-timed in the era of big sequence data. *Front Genet* 5:114
- Adami C (2004) Information theory in molecular biology. *Phys Life Rev* 1:3–22. doi:10.1016/j.plev.2004.01.002
- Adami C (2012) The use of information theory in evolutionary biology. *Ann N Y Acad Sci* 1256:49–65. doi:10.1111/j.1749-6632.2011.06422.x
- Altman RB (1993) Probabilistic structure calculations: a three-dimensional tRNA structure from sequence correlation data. *Proc Int Conf Intell Syst Mol Biol* 1:12–20
- Arnez JG, Moras D (1997) Structural and functional considerations of the aminoacylation reaction. *Trends Biochem Sci* 22:211–216
- Belousoff MJ, Davidovich C, Bashan A, Yonath A (2010) On the development towards the modern world: a plausible role of uncoded peptides in the RNA world. In: Ruiz-Mirazo K, Luisi PL (eds) *Origins of life and evolution of biospheres*. Springer, New York City, pp 415–419
- Bonnefond L, Giegé R, Rudinger-Thirion J (2005) Evolution of the tRNA^{Tyr}/TyrRS aminoacylation systems. *Biochimie*. 87:873–883
- Carter CW, Li L, Weinreb V et al (2014) The Rodin-Ohno hypothesis that two enzyme superfamilies descended from one ancestral gene: an unlikely scenario for the origins of translation that will not be dismissed. *Biol Direct* 9:11. doi:10.1186/1745-6150-9-11
- Cusack S (1997) Aminoacyl-tRNA synthetases. *Curr Opin Struct Biol* 7:881–889. doi:10.1016/S0959-440X(97)80161-3
- De Duve C (1988) The second genetic code. *Nature* 333:117–118. doi:10.1038/333117a0
- Durbin R, Eddy SR, Krogh A, Mitchison G (1998) *Biological sequence analysis*. Cambridge University Press, Cambridge
- Durbin R, Eddy S, Krogh A, Mitchison G (2002) *Biological sequence analysis*. Cambridge University Press, Probabilistic models of proteins and nucleic acids
- Eriani G, Delarue M, Poch O et al (1990) Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. *Lett to Nat* 347:203–206. doi:10.1038/346183a0
- Farias ST, Rêgo TG, José MV (2014a) Evolution of transfer RNA and the origin of the translation system. *Front Genet* 5:303. doi:10.3389/fgene.2014.00303
- Farias ST, Rêgo TG, José MV (2014b) Origin and evolution of the peptidyl transferase center from proto-tRNAs. *FEBS Open Bio* 4:175–178
- Farias ST, Rêgo TG, José MV (2016) tRNA core hypothesis for the transition from the RNA world to the ribonucleoprotein world. *Life (Basel, Switzerland)* 6:15. doi:10.3390/life6020015
- Farias ST, Rêgo TG, José MV (2017) Peptidyl transferase center and the emergence of the translation system. *Life (Basel, Switzerland)* 7:21. doi:10.3390/life7020021
- Giegé R, Sissler M, Florentz C (1998) Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Res* 26:5017–5035
- Goddard JP (1977) The structures and functions of transfer RNA. *Prog Biophys Mol Biol* 32:233–308. doi:10.1016/0079-6107(78)90021-4

- Guimarães RC, Moreira CHC, de Farias ST (2008) A self-referential model for the formation of the genetic code. *Theory Biosci* 127:249–270. doi:10.1007/s12064-008-0043-y
- Hou YM (2010) CCA addition to tRNA: implications for tRNA quality control. *IUBMB Life* 62:251–260
- Hou Y-M, Schimmel P (1988) A simple structural feature is a major determinant of the identity of a transfer RNA. *Nature* 333:140–145. doi:10.1038/333140a0
- José MV, Morgado ER, Guimarães RC et al (2014) Three-dimensional algebraic models of the tRNA code and 12 graphs for representing the amino acids. *Life (Basel, Switzerland)* 4:341–373. doi:10.3390/life4030341
- José MV, Zamudio GS, Farias ST (2016) Evolution of tRNAs was driven by entropic forces. Evolution of the protein synthesis machinery and its regulation. In: Jagus R, Hernández G (eds). Springer International Publishing Switzerland, pp 1–7. Book ID: 370644_1_En. doi: 10.1007/978-3-319-39468-8
- José MV, Zamudio GS, Morgado ER (2017) A unified model of the standard genetic code. *Royal Society Open Science* 4:160908. doi:10.1098/rsos160908
- Li W (1990) Mutual information versus correlation functions. *J Stat Phys* 60:823–837. doi:10.1007/BF01025996
- Lofffield RB (1972) The mechanism of Aminoacylation of transfer RNA. *Prog Nucleic Acid Res Mol Biol* 12: 87–128. doi:10.1016/S0079-6603(08)60660-1
- Meila M (2003) Comparing clusterings by the variation of information. *Learn theory Kernel Mach 16th Annu Conf Learn Theory 7th Kernel Work COLT/Kernel 2003*, Washington, DC, USA, August 24–27, 2003 Proc 173. doi: 10.1007/978-3-540-45167-9_14
- Meilã M (2007) Comparing clusterings-an information based distance. *J Multivar Anal* 98:873–895. doi:10.1016/j.jmva.2006.11.013
- Mukai T, Reynolds N, Crnković A, Söll D (2017) Bioinformatic analysis reveals archaeal tRNA^{Tyr} and tRNA^{Trp} identities in bacteria. *Life (Basel, Switzerland)* 7:8. doi:10.3390/life7010008
- Park SJ, Schimmel P (1988) Evidence for interaction of an aminoacyl transfer RNA synthetase with a region important for the identity of its cognate transfer RNA. *J Biol Chem* 15:16527–16530
- Ribas de Pouplana L, Turner RJ, Steer BA, Schimmel P (1998) Genetic code origins: tRNAs older than their synthetases? *Proc Natl Acad Sci U S A* 15:11295–11300
- Rodin SN, Ohno S (1997) Four primordial modes of tRNA-synthetase recognition, determined by the (G,C) operational code. *Proc Natl Acad Sci U S A* 94:5183–5188. doi:10.1073/pnas.94.10.5183
- Rodin SN, Rodin AS (2008) On the origin of the genetic code: signatures of its primordial complementarity in tRNAs and aminoacyl-tRNA synthetases. *Heredity (Edinb)* 100:341–355. doi:10.1038/sj.hdy.6801086
- Rodin SN, Rodin AS, Ohno S (1996) The presence of codon-anticodon pairs in the acceptor stem of tRNAs. *Proc Natl Acad Sci U S A* 93:4537–4542. doi:10.1073/pnas.93.10.4537
- Saks ME, Samposn JR, Abelson JN (1994) The transfer RNA identity problem: A search for rules. *Science* 263: 191–197
- Schimmel P (1995) An operational RNA code for amino acids and variations in critical nucleotide sequences in evolution. *J Mol Evol* 40:531–536. doi:10.1007/BF00166621
- Schimmel P, Giégé R, Moras D, Yokoyama S (1993) An operational RNA code for amino acids and possible relationship to genetic code. *Proc Natl Acad Sci U S A* 90:8763–8768. doi:10.1073/pnas.90.19.8763
- Shimizu M (1995) Specific aminoacylation of C4N hairpin RNAs with the cognate aminoacyl-adenylates in the presence of a dipeptide: origin of the genetic code. *J Biochem* 117:23–26
- Sundaharadas G, Katze JR, Söll D et al (1968) On the recognition of serine transfer RNA's specific for unrelated codons by the same seryl-tRNA synthetase. *Proc Natl Acad Sci U S A* 61:693–700
- Tamura K (2015) Origins and early evolution of the tRNA molecule. *Life* 5:1687–1699. doi:10.3390/life5041687
- Vargas-Rodríguez O, Musier-Forsyth K (2014) Structural biology: wobble puts RNA on target. *Nature* 510:480–481
- Wu XQ, Gross HJ (1993) The long extra arms of human tRNA(Ser)^{sec} and tRNA^{Ser} function as major identity elements for serylation in an orientation-dependent, but not sequence-specific manner. *Nucleic Acids Res* 21:5589–5594. doi:10.1093/nar/21.24.5589
- Yarus M (1982) Translational efficiency of transfer RNA's: uses of an extended anticodon. *Science* 218:646–52. doi:10.1126/science.6753149
- Zamudio GS, José MV (2017) On the uniqueness of the standard genetic code. *Life* 7:7. doi:10.3390/life7010007

Origins of Life & Evolution of the Biosphere is a copyright of Springer, 2018. All Rights Reserved.

Apéndice 6



Information theory unveils the evolution of tRNA identity elements in the three domains of life

Gabriel S. Zamudio¹ · Miryam Palacios-Pérez¹ · Marco V. José¹

Received: 21 April 2019 / Accepted: 3 September 2019 / Published online: 18 September 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract

We determined the identity elements of each tRNA isoacceptor for the three domains of life: Eubacteria, Archaea, and Eukarya. Our analyses encompass the most updated and curated available databases using an information theory approach. We obtained a collection of identity clusters for each of the isoacceptors of the 20 canonical amino acids for the three major domains of life. The identity clusters for all isoacceptors are compared within and among the three domains to determine their pattern of differentiation and to shed light on the evolution of the identity elements.

Keywords tRNA identity elements · Three domains of life · tRNA evolution · Information theory

Introduction

The translation machine comprises an ample set of molecules interacting in a complex biological network. At the center of such network, it stands out the transfer RNA (tRNA) interacting with different molecules, including the messenger RNA (mRNA), the aminoacyl-tRNA synthetases (aaRSs), and the ribosome. Molecular recognition is a process that entails a principle of memory. To maintain such a complex interaction scheme, tRNAs possess two recognition codes in its structure, the anticodons for the mRNA, and the one known as “operational code” for the aaRS (De Duve 1988; Ribas de Pouplana and Schimmel 2001). The tRNAs operational code conducts the correct pairing of a tRNA with its cognate aaRS which has been previously charged with its corresponding amino acid. The attachment of the amino acid from the aaRSs to a tRNA is through an esterification reaction on the 3' end of the tRNA (Arnez and Moras 1997). After a tRNA has been correctly aminoacylated, it conducts the correct translation of the mRNA into a peptide through the ribosome. There are 20 aaRSs (one for each

amino acid of the standard genetic code), and each aaRS can be paired to a set of tRNAs with different anticodons; this set of tRNAs is known as isoacceptors. The existence of only 20 different aaRSs makes the operational code non-degenerate (Eriani et al. 1990). The family of the aaRSs enzymes are divided into two subfamilies (Class I and Class II) (Eriani et al. 1990). Both the operational code and the anticodon code did not evolve independently (Zamudio and José 2018; de Farias et al. 2018) since the early emergence of the mini-helix structure of the tRNA 3.5 billion years ago (Tamura 2015).

Modern aaRSs do not, in some cases, directly read the tRNA's anticodon (Ribas de Pouplana and Schimmel 2001). Although there is no explicit recognition of the anticodon, coupled with the degeneracy of the standard genetic code, tRNAs are charged with the correct amino acid. This correct aminoacylation of a tRNA is made through the operational code that is comprised by a set of identity elements that echo the information of the anticodon on the rest of the tRNA structure. Different methods have been used to identify the location of the identity elements, including experimental analysis (Giegé et al. 1998), and different mathematical and computational approaches (Zamudio and José 2018; Mukai et al. 2017; Branciamore et al. 2018). The consensus is that the set of identity elements differs for each isoacceptor group. The identity elements have been proposed to participate in the recognition of tRNAs not only in the aminoacylation reaction but also in the tRNA–protein interaction network (Ardell 2010). We have previously determined

✉ Gabriel S. Zamudio
gazaso92@gmail.com

✉ Marco V. José
marcojose@biomedicas.unam.mx

¹ Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, C.P. 04510 Mexico City, CDMX, Mexico

the identity elements regardless of the three main domains of life (Zamudio and José 2018). Differences have been found in the tRNA recognition system of the three domains of life (Woese et al. 2000). This collection of differences has been proposed to set a barrier of interdomain horizontal gene transfer of the aaRS genes (Ardell 2010). Differences in the aaRSs are also reflected in the tRNAs identities of each domain. tRNA genes have been found with different configurations in the three domains, such as multiple introns, split and tri-split tRNAs, and permuted tRNAs (Fujishima and Kanai 2014).

A tRNA has a canonical length of 72 nucleotides and is divided into four main sectors: the acceptor stem, the D-arm, the anticodon arm, and the T-arm and a sector known as variable region. Each arm is composed by a loop and a stem. In the 5' to 3' sense, the acceptor stem is joined to the D-arm followed by the anticodon arm; next is the variable region which connects to the T-arm that returns to the acceptor stem, thus forming a closed structure (Fig. 1). The 3' end of the tRNA is capped with an extra nucleotide and a terminal CCA which is added posttranscriptionally (Tamaki et al. 2018; Hou 2010). The D-arm has uridines modified nucleotides to dihydrouridines (Motorin and Grosjean 2005), and the T-arm is also known as *T Ψ C*-arm due to the presence of thymidine, pseudouridine, and cytidine nucleotides. The variable region gets its name from the variable length it possesses.

We posed the question on how different the compendiums of the operational codes in the three main domains of life are. A hallmark of aaRSs is the exquisite specificity with which they select and aminoacylate only their cognate tRNA (Hendrickson 2001). Therefore, the discernment of the different operational codes becomes a central issue for a better understanding of the evolution of the translation system. In this work, the identity elements of the tRNAs are determined for each of the three domains of life using an information theoretical approach. We perform systematic analyses using the most updated and curated available databases (Jühling et al. 2009; Abe et al. 2014) (accessed August 2018). We derive a collection of identity clusters for each of the isoacceptors of the 20 canonical amino acids for the three major domains of life. The identity clusters for all the isoacceptors are compared within and among the three domains not only to determine their pattern of differentiation but also to gain insights on the origins and evolution of the identity elements and the translation system.

Data sources

Data of mature nuclear tRNA gene sequences were downloaded from Jühling et al. (2009) and Abe et al. (2014) (accessed August 2018) for the 20 canonical amino acids.

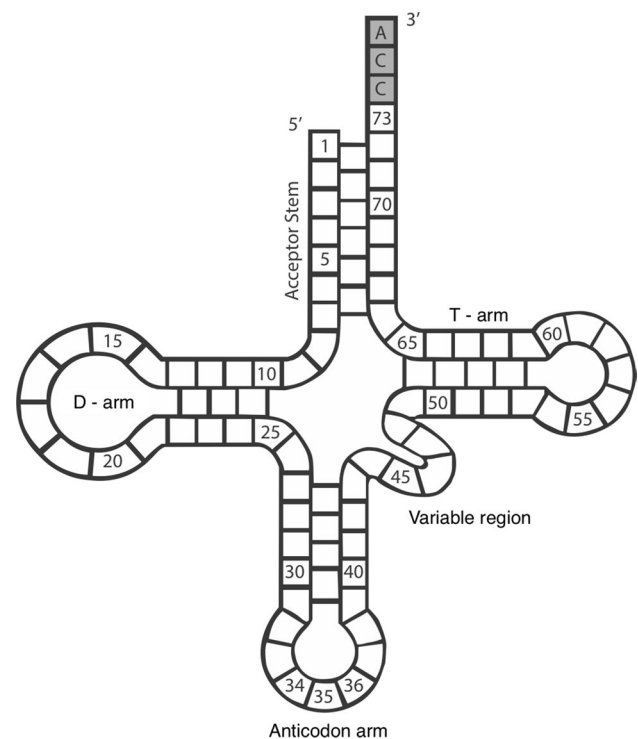


Fig. 1 Secondary structure of a tRNA with canonical length. The acceptor stem consists of the bases from 1 to 7 and from 66 to 72. The D-arm is constituted by bases 10 up to 25. The anticodon arm comprises the portion of bases from positions 27 to 43; the anticodon triplet is bases 34, 35, and 36. The variable region is made by bases from positions 44 to 48. The T-arm starts at base 49 and ends at position 65. The segment made by bases 8 and 9 connects the acceptor stem to the D-arm, while base 26 joins the D-arm with the anticodon arm. Base 73 and terminal CCA are added posttranscriptionally

The variable region of the sequences was removed to make the sequences comparable in length. The dataset was divided according to the three domains of life: Archaea, Eubacteria, and Eukarya, for the 20 tRNA isoacceptors. Two distinct databases were constructed from the tRNA sequences, due to the different lengths of the D-arm. In the first dataset, the D-loop was removed, while maintaining its respective stems (non-loop). For the second dataset, the length of the D-loop was determined per sequence and those whose length has more samples were considered (with loop). Duplicated sequences in each dataset were removed. The canonical length of the D-loop is eight nucleotides; however, in some cases, only one sequence exists, so the use of the canonical length was discarded in the analysis. The number of extra bases considered in the D-loop for the with-loop dataset is 1, 1, and 0, for Archaea, Eubacteria, and Eukarya, respectively. On the three domains, the isoacceptors sets for initiator methionine (ini-Met) and elongator methionine were differentiated according to the comments on the downloaded data. Isoacceptors sets with 15 or less sequences were not considered due to the low sample size.

Methods

Information theory quantifies the transmission and sharing of information through an information channel or between two messages. Given two messages, it is possible to determine the information shared by them. The variation of information is pseudometric (the distance between any two different elements can be zero), and it measures the information distance between two messages X and Y . The variation of information is given by $V(X, Y) = H(X) + H(Y) - 2I(X, Y)$, where $H(X)$ is the entropy of the random variable X and $I(X, Y)$ is the mutual information between the two random variables X and Y (Zamudio and José 2018; Meilă 2003). For the analysis of tRNA sequences, we define as a continuous random variable the nucleotides in a given site on the set of tRNA isoacceptors for each amino acid. This allows us to compute the information distance between two sites in the tRNA structure within the isoacceptor groups on each domain. If any two given sites on the tRNA have an information distance of 0, then the occurrences of bases in the two sites are completely predictable one from the other. The variation of information allows clustering sites according to the variation of information among them. On any given cluster, all the sites on it will have a variation of information less or equal to a given parameter d . By setting the parameter to some specific values, such as $d = 0$, the clusters are well defined, whereas with a positive parameter the clusters can in some cases be fuzzy, i.e., a site on a tRNA isoacceptor may belong to two or more identity clusters. The appearance of fuzzy clusters is due to the triangle inequality property of any metric function such as the variation of information. For every isoacceptor in each domain, the value of maximum variation of information d_{MAX} , which ensures that for all the values d_1 , such that $0 \leq d_1 \leq d_{\text{MAX}}$, the clusters inferred using the parameter d_1 are well defined, was found. The clusters of sites formed with this parameter d_{MAX} of maximum information distance within each other comprise the collection of identity elements of each isoacceptor.

The use of the value d_{MAX} for the definition of the identity clusters determines the maximum value that allows the construction of well-defined clusters on each tRNA isoacceptor, albeit the information distance between any two sites of the same identity cluster d' , may be lower than d_{MAX} , i.e., $d' \leq d_{\text{MAX}}$.

The supreme possible value of variation of information between any two clusters occurs when the variables X and Y are independent and uniformly distributed; in this case, the maximum value is $d = 4$, so in general the inequality $0 \leq d \leq 4$ holds.

Results

The variation of information for the three domains of life on the two datasets is computed. The positive value d_{MAX} for well-defined clusters was found, and several clusters of sites with an information distance lower than d_{MAX} , between the positions in each set are found for all the isoacceptors in the three domains for the with-loop dataset (Fig. 2) and for the non-loop dataset (“Appendix”). In Fig. 2, positions marked in red correspond to the set of sites related to the central anticodon base 35. Sets in different colors correspond to clusters of sites whose information distance is also lower than d_{MAX} . For the sake of clarity, pairs of sites corresponding to Watson–Crick pairs in the tRNA molecule that are at information distance lower than d_{MAX} are not marked. Positions marked in black correspond to the removed variable loop in both datasets, and the D-loop removed in the non-loop dataset. The numbering of the positions starts at the 5' end of the tRNA, and extra bases are marked with an asterisk (*), so that the second half of the D-stem begins at position 22. The value of the parameter d_{MAX} in Fig. 2 is normalized to the maximum value of 4 by using the value $d_{\text{MAX}}/4$. After the normalization, the distance values lie on the interval $0 \leq \frac{d_{\text{MAX}}}{4} \leq 1$. In some isoacceptors, the information distance, which defines the identity clusters, is almost zero $\approx 10^{-16}$, which shows that the sites on the identity clusters are highly conserved on the isoacceptor with the exception of a tiny amount of rare mutations present in the dataset. For the rest of the isoacceptors, the higher values of information distance show that the sites are not highly conserved for a fixed nucleotide; instead, the nucleotides at the sites of an identity cluster follow a changing pattern which is predictable and thus conserved.

For archaeal tRNAs, the first four positions of the terminal side of the acceptor stem are present in clusters of sites related to the anticodon or with other positions in both datasets. Neither the bacterial nor the eukaryal tRNAs exhibit this pattern, albeit there are some exceptions. Positions 8 and 9 bridge the acceptor stem with the D-arm. Position 8 has a stronger presence in the identity clusters of all isoacceptors than position 9; in the cases where position 8 is associated with an identity cluster, it is usually associated with cluster containing the anticodon. For bacterial tRNAs, position 8 is present on some identity cluster for seven isoacceptors (Ala, Arg, Gln, Gly, Thr, Trp, and Val) in the with-loop dataset, from which Ala, Gly, Thr, and Val are amino acids found in Miller’s experiment (1953, 1957, 1974). For the Eukarya domain, position 8 is found in eight isoacceptors on both datasets. For the Archaea domain, position 8 is constantly absent from the identity clusters on the non-loop dataset with the

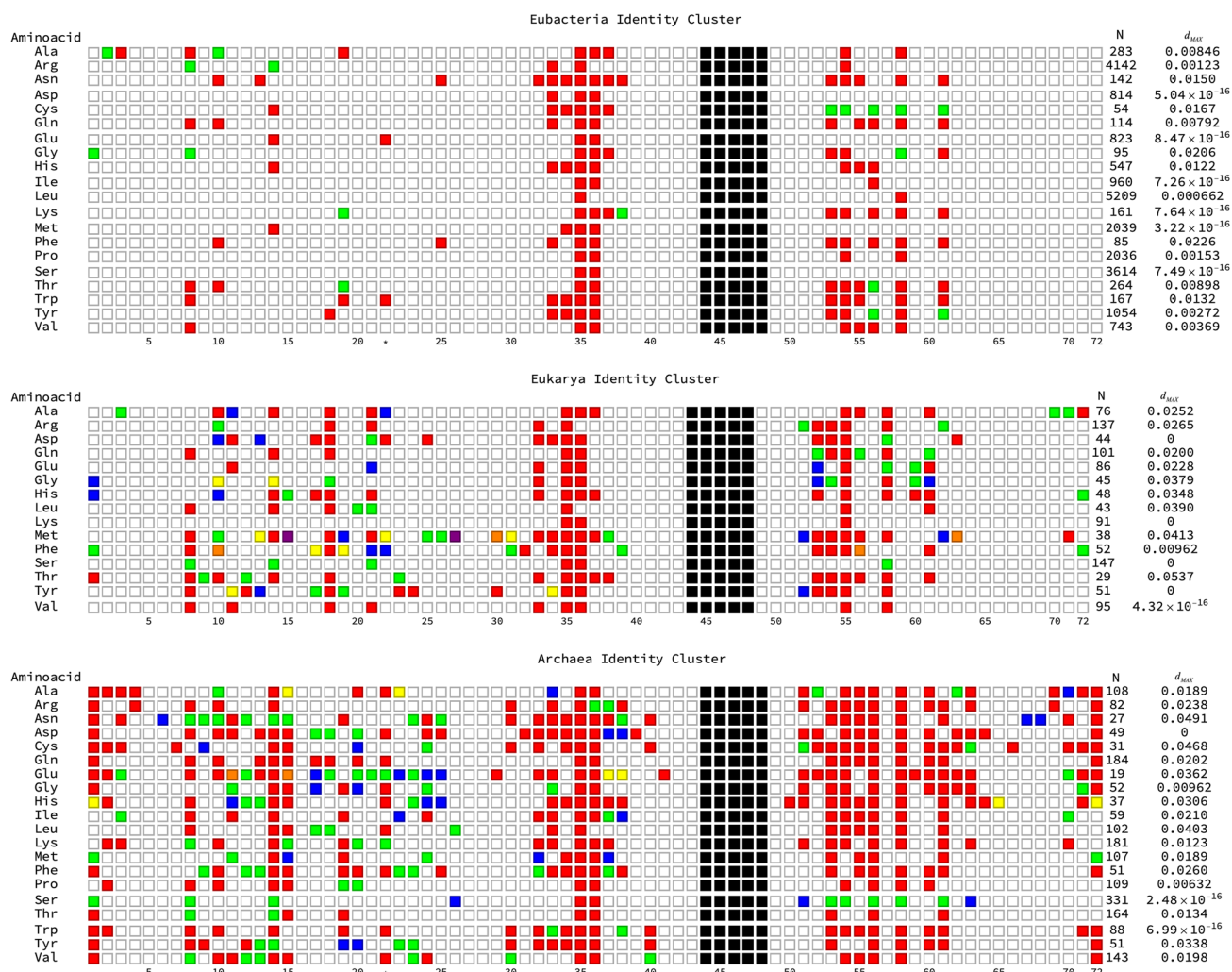


Fig. 2 Identity clusters derived from the with-loop dataset. Extra sites in the D-loop are marked with asterisk (*) after the end of the canonical D-loop at position 21. Sites removed of the variable region are in black. On each isoacceptor, the different colors (red, blue, green, yellow orange, purple) refer to the disjoint identity clusters grouping the sites of the each tRNA isoacceptor. The principal cluster containing the central anticodon base is colored in red. The value N stands

for the sample size of each isoacceptor group. The d_{MAX} value on the isoacceptors is the information distance that determines the identity clusters for each isoacceptor. The d_{MAX} value shown has been normalized to the maximum number of information distance of 4, as $\frac{d_{MAX}}{4}$. A similar figure derived for the non-loop dataset is provided in “Appendix” (color figure online)

exceptions of Gln and Pro (on which it is associated with the anticodon), and also for Asn and Glu. In contrast, position 8 is widely present on some identity clusters on the with-loop dataset. The D-stem is composed by positions 10–13 and 22–25. For the non-loop dataset of Eubacteria and Eukarya, there are some cases with clusters conformed only by bases forming Watson–Crick pairings, and only position 10 of Phe in Eukarya is part of a wider cluster. An ample number of identity elements appear in the D-stem for the archaeal non-loop dataset. In the with-loop dataset, bacterial tRNAs display a number of identity elements along the D-arm which increases for Eukarya and Archaea. For the archaeal D-loop, there is a constant occurrence of position 14 as being part of an identity cluster. This

occurrence is less present in Eukarya, while for Eubacteria position 14 is contained in an identity cluster in only five isoacceptors. The anticodon arm is composed of bases 27–43, with position 26 connecting the D-arm with the anticodon arm. In the three domains, for both datasets, the first three base pairs of the anticodon stem, to wit, 27–43, 28–42, and 29–41, in general do not belong to any wide identity cluster, with two exceptions in the with-loop dataset. Such exceptions are the isoacceptor for Glu in the Archaea whose domain has the pair 29–41 related to the anticodon and for Met in Eukarya where position 27 is related to position 15. The anticodon loop in all domains possesses bases that are part of the identity cluster associated with the central position 35 of the anticodon. Note

that in some cases, the loop bases surrounding the anticodon triplet are part of an identity cluster which is not associated with the anticodon. The T-arm is composed of the positions 49–65 in the canonical structure. The tRNAs for Archaea present a wider spectrum of identity elements in T-arm for both datasets than the other two domains. The clusters of identity elements for the non-loop Eukarya tRNAs have a wider presence in the T-stem than in the corresponding bacterial tRNAs; meanwhile, on the with-loop dataset, the domains of Eukarya and Eubacteria have an average of equal identity elements across the T-arm. Position 57, which is the central position of the T-loop, does not belong to any identity cluster in any domain, while the surrounding bases are contained in a cluster in most cases.

Discussion

In this work, we calculated the identity elements of the 20 tRNA canonic isoacceptors for Eubacteria, Archaea, and Eukarya domains using a metric from information theory. Current datasets contain tRNA gene sequences with different lengths of the D-loop; hence, sequences were analyzed with and without the D-loop. An updated catalog of the identity elements for each of the main domains of life is presented. The purpose of using two datasets was to determine whether the length of the D-loop is associated with the identity elements of the operational code in tRNAs. Subtle differences in the identity elements appeared in the D-stems that show that, in some cases, the D-loop influences the operational code. The variable region was deleted in order to make the tRNA sequences generally equal in length. Different identity clusters are found for each isoacceptor on the three domains. In the three domains, the with-loop isoacceptors present an increased number of identity elements in the D-stem than their corresponding counterparts of the non-loop dataset. A general pattern appears that resembles the phylogenetic tree of the three domains. Most of the isoacceptors in the three domains present identity elements in the T-arm, while tRNAs from Eukarya and Archaea possess a higher number of identity elements in the D-arm when compared to Eubacteria. Finally, archaeal isoacceptors show identity elements in the acceptor stem, which are absent in the other two domains. There is a high bias in repositories of bacterial tRNA sequences; Eukarya and Archaea have sample sizes of the same magnitude. The use of a positive value d as the parameter of the variation of information allows for the analysis of not completely strict similarity patterns. The variation of information between two sites, as information metric, gives the information necessary to discern the value of one parameter given the knowledge of the other. A variation of information of 0 between two sites results in the need

of no more information to determine the value of one site from previously knowing the other. The clustering parameter, d_{MAX} , determines the magnitude of the variability for all the identity clusters in any given isoacceptor.

The widely established base pair G3:U70 determinant in tRNA^{Ala} (Ribas de Pouplana and Schimmel 2001; Hou and Schimmel 1988; McClain and Foss 1988; Chong et al. 2018) is part of an identity cluster in the three domains. In bacterial tRNAs^{Ala}, base 3 is on the same cluster as the anticodon; for the Eukarya domain, a cluster conformed by bases 3, 70, and 71 is formed; in Archaea, base 3 is related to the anticodon central position, while base 70 is related to base 33 which delimits the anticodon triplet. It has been reported that eukaryal AlaRS has gained functionality by mischarging non-cognate tRNAs due to the recognition of the pair G4:U69 (Sun et al. 2016). Archaeal AlaRS possesses the same mechanisms for detecting the G4:U69 base pair in non-cognate tRNAs (Sun et al. 2016); this base pair is present in the anticodon identity cluster for tRNAs^{Ala} in the Archaea domain. Mechanisms for correction of the mischarging of tRNA^{Thr} with alanine by AlaRS in kingdom Animalia have been described (Kuncha et al. 2018). The G:U wobble base pair is a fundamental unit of RNA secondary structure in all three phylogenetic domains (Varani and McClain 2000).

Differences with the identity clusters found previously arise (Zamudio and José 2018). This is a consequence of using the three domains in the same dataset and restricted the analyses to sequences with the D-loop of eight nucleotides, which is the canonic length, coupled with the use of an information distance that allows variability.

A distinctive feature of tRNA^{His} is an extra 5' nucleotide that is usually a guanylate at position G:-1 (Wang et al. 2007). This nucleotide is added posttranscriptionally, and therefore, it was not included in our gene analysis.

Some bases in the identity clusters correspond to positions associated with posttranscriptional modifications which have been reported to be either universal in the three domains of life (Jühling et al. 2009), generally present in two domains, or are domain specific (Motorin and Grosjean 2005; Lorenz et al. 2017). Modifications of nucleobases from posttranscriptional modifications enhance the stability of tRNAs and improve its interaction with other molecules involved in translation, such as aaRS, translation factors, or the mRNA (Motorin and Grosjean 2005).

The D-arm receives its name as it contains the modified base dihydrouridine (Lorenz et al. 2017), which is the result of adding two hydrogen atoms to a uridine nucleoside. The D-arm provides structural stability to the tRNA and avoids its premature dissociation from the ribosome (Smith and Yarus 1989). Such a degree of interaction between the D-arm with aaRS is more notorious in bacterial and eukaryal tRNAs, whereas such arm does not seem imperative for

Archaea (Tamaki et al. 2018), which could help to explain the identity clusters and its dendrograms. However, the D-arm confers a more precise differentiation between each other tRNA. On the D-loop, the positions associated with the anticodon are not the ones which are modified to dihydrouridines, which provide flexibility to this region (Motorin and Grosjean 2005). One of these bases is position 14 which belongs to an identity cluster in some bacterial tRNAs, half of the eukaryal isoacceptors, and it is a general property of archaeal tRNAs and is not modified to dihydrouridine. Positions of tRNA anti-determinant bases, i.e., positions on which the presence of a specific base disassociates the recognition of the tRNA with its corresponding aaRS, are not generally discernible by our methodology. Such is the case of C:34 for bacterial tRNA^{Ile} which is not present on any identity cluster; the opposite example is position 10 for eukaryal tRNA^{Phe}. This position has been reported as an important base for the recognition with its corresponding aaRS on yeast (Motorin and Grosjean 2005). In addition, this position arises as an identity element associated with position 56 on the T-loop. Positions 10 and 56 are on opposite sides of the tRNA on the cloverleaf representation and are arranged on opposite sides at the corners in the L-shaped tertiary structure. The nucleotides on positions 8 and 10 undergo modifications to thiouridine for photon protection (Motorin and Grosjean 2005) and N2-methylguanosine for proper folding (Lorenz et al. 2017), respectively. Position 8 is related to the anticodon on some bacterial tRNAs, whereas position 10 is an identity element for some eukaryal and archaeal tRNAs. Our results suggest intricate relationships between the positions related to the structural properties and preservation and the identity elements of the tRNAs. Giegé stated that identity elements are rare in D-arm; however, some determinants in the tRNAs for certain amino acids have been characterized (Hendrickson 2001). Herein, we report that identity elements in the D-loop of Archaea and Eukarya are ubiquitous. For example, position 20 in bacteria does not appear in any identity cluster, whereas in Archaea this position is fixed in the 20 amino acids and in Eukarya appears in eight amino acids. It has been suggested the existence of multiple sets of identity elements for each isoacceptor (Giegé et al. 1998), and this has been corroborated (Zamudio and José 2018).

The positions in the anticodon loop adjacent to the anticodon triplet are subject to modifications of the nucleosides in order to ensure accuracy and efficiency in translation (Motorin and Grosjean 2005); this property is reflected in the fact that in several isoacceptors the positions around the anticodon are in the same cluster as the anticodon central base. Likewise, according to reported tRNA molecule sequences (Jühling et al. 2009), position 57, in the

canonical structure length, remains unmodified in the vast majority of organisms and this base is consistently unassociated with any identity cluster in the three domains of life. Positions 55 and 58 are modified to pseudouridine and 1-methyladenosine, respectively, in the three domains of life, and both positions are associated with an identity cluster, generally the cluster for the anticodon, on most tRNAs of the three domains. For archaeal tRNAs, positions 55 and 58 are associated with base 56, and these three positions could indicate major recognition sites between the aaRSs and its cognate tRNA. On the D-loop, the positions associated with the anticodon are not the ones which are modified to dihydrouridines, which provide flexibility to this region (Motorin and Grosjean 2005). One of these bases is position 14 which belongs to an identity cluster in some bacterial tRNAs, half of the eukaryal isoacceptors, and it is a general property of archaeal tRNAs and is not modified to dihydrouridine. Positions of tRNA anti-determinant bases, i.e., positions on which the presence of a specific base disassociates the recognition of the tRNA with its corresponding aaRSs, are not generally discernible by our methodology. Such is the case of C:34 for bacterial tRNA^{Ile} which is no present on any identity cluster; the opposite example is position 10 for eukaryal tRNA^{Phe}; this position has been reported as an important base for the recognition with its corresponding aaRSs on yeast (Motorin and Grosjean 2005). This position arises as an identity element associated with position 56 on the T-loop. Positions 10 and 56 are on opposite sides of the tRNA on the cloverleaf representation and are arranged on opposite sides of the corner in the L-shaped tertiary structure. The nucleotides on positions 8 and 10 undergo modifications to thiouridine for photon protection (Motorin and Grosjean 2005) and N2-methylguanosine for proper folding (Lorenz et al. 2017), respectively. Position 8 is related to the anticodon on some bacterial tRNAs, whereas position 10 is an identity element for some eukaryal and archaeal tRNAs. Our results suggest intricate relationships between the positions related to the structural properties and preservation and the identity elements of the tRNAs.

We remark that when comparing the information theory approach with a conservation analysis, the information theory extends the results from a conservation analysis. If a base is fully conserved in a tRNA sequence, it will also be shown by the information theory approach because its variation of information is zero. If a base is less conserved, then its variation of information will be greater than zero. The less conserved a base is, relative to the others, the more will increase its variation of information.

The anticodon and acceptor arms are fully conserved but that is not the case with the variable arm. Whereas in middle range of conservation it would be the D and T

arms, D-arm confers the highest interaction with an aaRSs of the two (Tamaki et al. 2018). In agreement with the structural analysis by Tamaki et al. (2018), we found that positions 8, 10, 14, 19, 33, 53, 54, 55, 56, 58, and 61 are mostly conserved in the three domains of life.

Identity elements have functions beyond the correct interaction of a tRNA with its corresponding aaRS. The operational code also plays a role in guiding the correct folding of the tRNA to its tertiary structure. This property is manifested with the cluster of bases 10 and 56 that is present in Phe of Eukarya. Bases 10 and 56 belong to the D-loop and the T-loop, respectively, and they come into contact when the tRNA folds into its tertiary structure. In contrast to Giegé et al. (1998), we found tRNA identity nucleotides in the anticodon loop of bacterial tRNA^{Leu}, tRNA^{Ser}, and tRNA^{Ala}.

The D-arm receives its name as it contains the modified base dihydrouridine (Lorenz et al. 2017), which is the result of adding two hydrogen atoms to a uridine nucleoside. The D-arm provides structural stability to the tRNA and avoids its premature dissociation from the ribosome (Smith and Yarus 1989). Such a degree of interaction between the D-arm with aaRS is more notorious in bacterial and eukaryal tRNAs, whereas such arm does not seem imperative for Archaea (Tamaki et al. 2018), which could help to explain the identity clusters and its dendrograms. However, the D-arm confers a more precise differentiation between each other tRNA. Giegé stated that identity elements are rare in D-arm; however, some determinants in the tRNAs for certain amino acids have been characterized (Hendrickson 2001). Herein, we report that identity elements in the D-loop of Archaea and Eukarya are ubiquitous. For example, position 20 in bacteria does not appear in any identity cluster, whereas in Archaea this position is fixed in the 20 amino acids and in Eukarya appears in eight amino acids. It has been suggested the existence of multiple sets of identity elements for each isoacceptor (Giegé et al. 1998), and this has been corroborated (Zamudio and José 2018). Each tRNA molecule is usually composed by the D-arm, the anticodon arm, the variable region, the T-arm, and the acceptor stem with the addition of the terminal CCA added posttranscriptionally (Hou 2010). There are some few D-armless organisms and even one species described with only the anticodon and acceptor regions (Fujishima and Kanai 2014). The high number of identity elements in Archaea may provide robustness to the recognition system, given that several types of disruption of tRNA molecules have been observed, while still maintaining its functionality (Fujishima and Kanai 2014).

Carter and Wills (2018) have performed regression analyses of different qualitative features of bacterial

tRNAs, revealing that the acceptor stem of bacterial tRNAs retains an ancient operational code based on thermodynamic attributes, which are recognized by the corresponding aaRS. Our work provides a way to detect identity elements based on the mutual information of the anticodon with respect to other sites throughout the tRNA molecule, i.e., we look how the information contained in the anticodon could be reflected in other positions, which can be recognized by the aaRS not “seeing” the anticodon; our work is independent of structural considerations, which seems to be of high importance for bacterial tRNAs (Carter and Wills 2018), for which our approach does not detect many identity elements, in contrast to more complex or extremophile organisms such as Archaea, that usually have minimal tRNAs. Therefore, the abundance of identity elements may be necessary to guarantee the correct aminoacylation.

A wide number of functions of uncharged tRNAs outside the translation framework have been recently found. These roles include gene regulation, degradation, and cellular apoptosis (Raina and Ibba 2014). The identity elements in each isoacceptor could be related to the interaction of tRNAs in different biological networks.

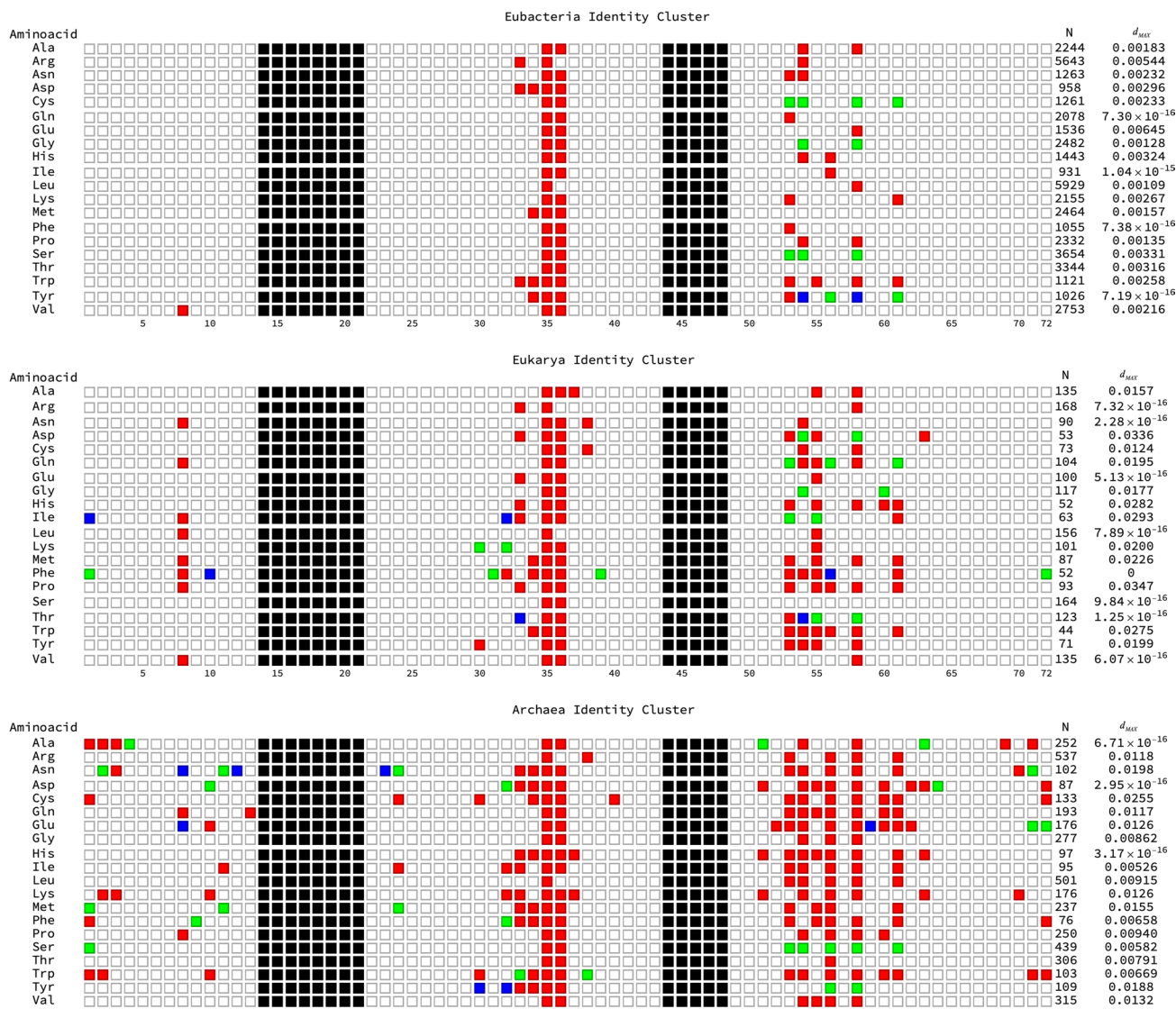
We were able to use the information theory to capture the evolution of the operational code; particularly, we tracked the information associated with the anticodon throughout the whole tRNA structure, which could help to explain more accurately how the aaRS can charge the correct amino acid without “seeing” the anticodon. This kind of approach would certainly be robust with the repository of more and diverse archaeal tRNAs. Additionally, the combination of different methods could improve the elucidation of the operational code and its evolution.

Acknowledgements We thank Juan R. Bobadilla for technical computer support. We thank the anonymous reviewer for their helpful criticisms and suggestions. We thank Adhemar Liquitaya-Montiel for helpful discussions at the beginning of this work.

Authors' contributions GSZ and MVJ conceived and designed the experiments; GSZ performed the experiments and analyzed the identity clusters; MPP analyzed tRNA sites for posttranscriptional modifications; GSZ, MPP, and MVJ analyzed the data; GSZ and MVJ wrote the paper.

Funding GSZ is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM), and received a doctoral fellowship from CONACYT (Number: 737920). MP is a doctoral student from Programa de Doctorado en Ciencias Biomédicas (PDCB), Universidad Nacional Autónoma de México (UNAM), and she receives the fellowship 694877 from CONACYT. MVJ was funded by Dirección General de Asuntos del Personal Académico (DGAPA), Universidad Nacional Autónoma de México, UNAM (PAPIIT-IN201019).

Appendix



References

- Abe T, Inokuchi H, Yamada Y, Muto A, Iwasaki Y, Ikemura T (2014) TRNADB-CE: tRNA gene database well-timed in the era of big sequence data. *Front Genet* 5:114
- Ardell DH (2010) Computational analysis of tRNA identity. *FEBS Lett* 584:325–333
- Arnez JG, Moras D (1997) Structural and functional considerations of the aminoacylation reaction. *Trends Biochem Sci* 22:211–216
- Branciamore S, Gogoshin G, Di Giulio M, Rodin A (2018) Intrinsic properties of tRNA molecules as deciphered via bayesian network and distribution divergence analysis. *Life* 8:5
- Carter CW, Wills PR (2018) Hierarchical groove discrimination by Class I and II aminoacyl-tRNA synthetases reveals a palimpsest of the operational RNA code in the tRNA acceptor-stem bases. *Nucleic Acids Res* 46:9667–9683
- Chong YE, Guo M, Yang X-L, Kuhle B, Naganuma M, Sekine S-I, Yokoyama S, Schimmel P (2018) Distinct ways of G: U recognition by conserved tRNA binding motifs. *Proc Natl Acad Sci* 115:7527–7532
- De Duve C (1988) The second genetic code. *Nature* 333:117–118
- de Farias ST, Antonino D, Rêgo TG, José MV (2018) Structural evolution of Glycyl-tRNA synthetases alpha subunit and its implication in the initial organization of the decoding system. *Prog Biophys Mol Biol* 30:1e8
- Eriani G, Delarue M, Poch O, Gangloff J, Moras D (1990) Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. *Nature* 347:203–206
- Fujishima K, Kanai A (2014) tRNA gene diversity in the three domains of life. *Front Genet* 5:142

- Giegé R, Sissler M, Florentz C (1998) Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Res* 26:5017–5035
- Hendrickson TL (2001) Recognizing the D-loop of transfer RNAs. *Proc Natl Acad Sci* 98:13473–13475
- Hou Y-M (2010) CCA addition to tRNA: implications for tRNA quality control. *IUBMB Life* 62:251–260
- Hou Y-M, Schimmel P (1988) A simple structural feature is a major determinant of the identity of a transfer RNA. *Nature* 333:140–145
- Jühling F, Mörl M, Hartmann RK, Sprinzl M, Stadler PF, Pütz J (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res* 37:D159–D162
- Kuncha SK, Mazeed M, Singh R, Kattula B, Routh SB, Sankaranarayanan R (2018) A chiral selectivity relaxed paralog of DTD for proofreading tRNA mischarging in Animalia. *Nat Commun* 9:511
- Lorenz C, Lünse C, Mörl M (2017) tRNA modifications: impact on structure and thermal adaptation. *Biomolecules* 7:35
- McClain WH, Foss K (1988) Changing the identity of a tRNA by introducing a G-U wobble pair near the 3' acceptor end. *Science* (80-) 240:793–796
- Meilä M (2003) Comparing clusterings by the variation of information. In: *Proceedings of learning theory and kernel machines: 16th annual conference on learning theory and 7th kernel workshop, COLT/Kernel 2003, Washington, DC, USA, August 24–27, 2003*. Springer, Berlin, pp 173–187
- Miller SL (1953) A production of amino acids under possible primitive earth conditions. *Science* (80-) 117:528–529
- Miller SL (1957) The mechanism of synthesis of amino acids by electric discharges. *Biochim Biophys Acta* 23:480–489
- Miller SL, Orgel LE (1974) *The origins of life on the earth*. Prentice-Hall, Upper Saddle River
- Motorin Y, Grosjean H (2005) tRNA modification. In: *Encyclopedia of life sciences*. Wiley. <https://doi.org/10.1038/npgs.els0003866>
- Mukai T, Reynolds N, Crnković A, Söll D (2017) Bioinformatic analysis reveals archaeal tRNA^{Tyr} and tRNA^{Trp} identities in bacteria. *Life* 7:8
- Raina M, Ibba M (2014) tRNAs as regulators of biological processes. *Front Genet* 5:171
- Ribas de Pouplana L, Schimmel P (2001) Operational RNA code for amino acids in relation to genetic code in evolution. *J Biol Chem* 276:6881–6884
- Smith D, Yarus M (1989) Transfer RNA structure and coding specificity. I. Evidence that a D-arm mutation reduces tRNA dissociation from the ribosome. *J Mol Biol* 206:489–501
- Sun L, Gomes AC, He W, Zhou H, Wang X, Pan DW, Schimmel P, Pan T, Yang XL (2016) Evolutionary gain of alanine mischarging to noncognate tRNAs with a G4:U69 base pair. *J Am Chem Soc* 138:12948–12955
- Tamaki S, Tomita M, Suzuki H, Kanai A (2018) Systematic analysis of the binding surfaces between tRNAs and their respective aminoacyl tRNA synthetase based on structural and evolutionary data. *Front Genet* 8:227
- Tamura K (2015) Origins and early evolution of the tRNA molecule. *Life* 5:1687–1699
- Varani G, McClain WH (2000) The G-U wobble base pair. *EMBO Rep* 1:18–23
- Wang C, Sobral B, Williams K (2007) Loss of a universal tRNA feature. *J Bacteriol* 189:1954–1962
- Woese CR, Olsen GJ, Ibba M, Soll D (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol Mol Biol Rev* 64:202–236
- Zamudio GS, José MV (2018) Identity elements of tRNA as derived from information analysis. *Orig Life Evol Biosph* 48:73–81

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Apéndice 7



A neutral evolution test derived from a theoretical amino acid substitution model

Gabriel S. Zamudio^a, Francisco Prosdocimi^b, Sávio Torres de Farias^c, Marco V. José^{a,*}

^aTheoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, CDMX, C.P. Ciudad de México 04510, Mexico

^bLaboratório de Biologia Teórica e de Sistemas, Instituto de Bioquímica Médica Leopoldo de Meis, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

^cLaboratório de Genética Evolutiva Paulo Leminsk, Departamento de Biologia Molecular, Universidade Federal da Paraíba, João Pessoa, Paraíba, Brazil

ARTICLE INFO

Article history:

Received 31 October 2018

Revised 14 January 2019

Accepted 28 January 2019

Available online 31 January 2019

Keywords:

Protein evolution

Positive selection

Negative selection

Neutral mutations

Neutrality test

ABSTRACT

A neutral evolution model that explicitly considers codons, amino acids, and the degeneracy of the genetic code is developed. The model is built from nucleotides up to amino acids, and it represents a refinement of the neutral theory of molecular evolution. The model is based on a stochastic process that leads to a stationary probability distribution of amino acids. The latter is used as a neutral test of evolution. We provide some examples for assessing the neutrality test for a small set of protein sequences. The Jukes-Cantor model is generalized to deal with amino acids and it is compared with our neutral model, along with the empirical BLOSUM62 substitution model. The neutral test provides a baseline to which the evolution of any protein can be analyzed, and it clearly helps in discerning putative amino acids with unexpected frequencies that might be under positive or negative selection. Our model and neutral test are as universal as the standard genetic code.

© 2019 Elsevier Ltd. All rights reserved.

1. Introduction

Understanding the processes that result in variations along protein evolution is central in molecular evolution and structural biology (Pál et al., 2006). Substitution models (SM) are widely used to portray the best approximation for the evolutionary path of genes and proteins. The first SM were based in (i) theoretical models for nucleotide changes (Jukes and Cantor, 1969) or (ii) the alignment of orthologous proteins from organisms (Dayhoff et al., 1978). Eventually, SM became more complex in order to include different physicochemical and biological factors such as the proportion of transitions and transversions (Kimura, 1979), and a plethora of hypotheses (Arenas, 2015). Besides to the fact that the application of such models and their interpretation is complex, most common software are not capable to directly incorporate SM with sophisticated mathematical and biological properties, such as site-dependence (Arenas, 2015). According to the evolutionary theory, mutational processes are considered to be neutral (Ng and Henikoff, 2006), as proposed by Kimura's neutral theory of evolution (Kimura, 1979). The neutral theory states that most changes at a molecular level are not driven by selection pressures

but by random fixation and neutral drift (Kimura, 1991). Neutrality was further extended to incorporate mutations that are slightly driven by selective pressure or nearly neutral processes (Ohta, 1973). Herein, we propose a new amino acid SM based on a mathematical model of the standard genetic code (SGC). This model represents the codons of the SGC as the vertices of a six-dimensional (6D) cube (José et al., 2017, 2012, 2007), and the edges represent point mutations in a codon. This codon-based model has been translated into its phenotypic graph representation (José et al., 2014), that allows the creation of an amino acid SM. The amino acid probability substitutions in this SM, are solely related to the degeneracy of the SGC, producing non-uniform probabilities.

With this substitution model, we present here a neutrality test based on the stationary distribution of the SM. The neutrality test allows a *bona fide* representation for the amount of amino acids expected exclusively by neutral evolution. By comparisons of the amount of amino acids actually observed in proteins, our SM allows the identification of amino acids that deviate from neutral expectations, permitting the identification of putative residues of positive or negative selection in any given protein.

2. Material and methods

The SGC has been mathematically modeled into a structure equivalent to a 6D-cube using group theory (José et al., 2007; Zamudio and José, 2017). In this model, the vertices of the cube

* Corresponding author.

E-mail addresses: marcojose@biomedicas.unam.mx, marcojose@cicc.unam.mx (M.V. José).

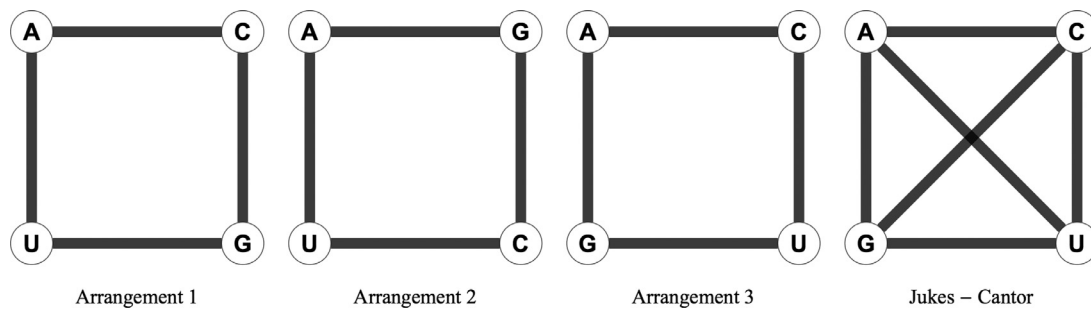


Fig. 1. Three possible arrangements of the four nucleotides as the vertices of a square that are not symmetrically equivalent (first 3 squares), representation of the JC-substitution model where all mutations are equally likely (right square with diagonals).

represent the codons, and the edges join codons that differ by one nucleotide given the possible vicinities of the nucleotides arranged in a square (note the first 3 squares of Fig. 1) (José et al., 2017; Zamudio and José, 2017). There are three possible ways to index the 64 codons into the vertices of a 6D cube (José et al., 2017). These ways are given by the three possible arrangements of four nucleotides as the vertices of a square that are not symmetrically equivalent, i.e. there is no symmetrical transformation that can transform one arrangement into another (note the first 3 squares of Fig. 1). The 6D-model has been further transformed into its amino acid phenotypic graph (APG) representation (José et al., 2015, 2014). In this APG, the vertices represent the amino acids; and two amino acids (aa_1 and aa_2) are joined by an edge; at least one codon that encodes aa_1 is at one edge distance to at least one codon that encodes aa_2 in the 6D-model of the SGC. The 6D representation of the genetic code is a 6D hypercube, which means that any codon has 6 other codons to which it is adjacent. For a given amino acid, the set of adjacencies of its codons represent the possible changes the amino acid can undergo with a single nucleotide mutation. If uniform weights are assigned to each substitution (the edges of the 6D-cube), the APG edges turn into weighted edges by adding the number of one-step substitution of all codons encoding a given amino acid that are neighbors (at one-edge distance) to codons encoding other amino acids. For example, if there are two edges joining codons for Ala and Ser in the 6D-cube, then, in the APG, the edge that join Ala and Ser will have a weight of two. Edges between codons for the same amino acid are counted twice as the substitution between these codons can happen in both ways, allowing the existence of weighted loops that account for the degeneracy of the SGC. By removing the stop codons of the 6D-cube, the adjacency matrix of the APG will contain the weights between all the edges of the 20 amino acids. The normalization by rows of this adjacency matrix leads to a probability transition matrix of a stochastic process (Bressloff, 2014), which is non-symmetric and whose states are the amino acids. The values from the stationary distribution determine the probability of finding each amino acid in any single position of the protein. Adding the Markov property (Bressloff, 2014) to the stochastic process results in a discrete time stochastic process with no memory. The limiting stationary distribution of a probability transition matrix is interpreted as the average time that any state is present in the long run (Sigman, 1995). As positive control, we use the neutral Jukes-Cantor model (JC-model) for nucleotides where all mutations are equally likely (Jukes and Cantor, 1969). This means that all possible nucleotide changes can be achieved in a one-step transformation (note the right square with diagonals of Fig. 1). A similar procedure to construct the geometric representation of the codons

is also considered. This yields a similar 6D-cube, although it differs from the previous ones as some of its diagonals appear. The same procedure can be applied to derive its corresponding phenotypic graph, stochastic process, and its limiting stationary distribution. Therefore, the original nucleotide-based JC-model (Jukes and Cantor, 1969) is extended to deal explicitly with amino acids and with the degeneracy of the SGC. As a negative control, we use the transition matrix from BLOSUM62 (Henikoff and Henikoff, 1992). BLOSUM62 matrices are derived from highly conserved regions of protein families by counting the amino acids changes occurring in the alignments of the sequences of these regions. A comparison with the BLOSUM62 matrix is done by retrieving its conditional probabilities using the method, computational script, and data available in (Eddy, 2004). As these conditional probabilities determine the probability of a state given another state in the alignment for constructing the BLOSUM62 matrix, the matrix of conditional probabilities determines a stochastic process, from which its limiting stationary distribution is calculated.

3. Results

The only three possible APGs, from the nucleotides arranged in a square, yield three different stochastic processes. These processes differ by their probability transition matrix; the average of these transition matrices is a stochastic process that equally considers and weights all the possible transitions at nucleotide level, with a slight bias toward maintaining transitions since two out of the three squares arrange purines and pyrimidines as neighbors, and the third one intercalates the chemical types, i.e., a mutation in one step interchanges a purine with a pyrimidine and vice versa.

The limiting distribution (or probability stationary distribution), of the average transition matrix is a neutral control of the changes present in a protein history at amino acid level. This neutral control is to be compared to the stationary distribution from a matrix of accepted mutations derived from actual protein sequences. Our neutrality test, therefore, compares the individual components of a stationary distribution derived from protein sequences with the neutral control. Herein, we dubbed our test the amino acid neutrality test (ANT). The neutral control indicates the probability of each amino acid to be present in a hypothetical protein that has been target of random mutations at the nucleotide level. Thus, ANT considers all positions as equally likely to change, with the sole constraint that they obey the degeneracy of the genetic code. If a component from the stationary distribution, obtained from protein sequences, has a greater value than its corresponding component at the neutral control, it will be interpreted as positive selection, and lower values will be considered as negative selection. The values that lie at or close to the neutral control will be interpreted as consistent with the neutral mutation-random drift hypothesis of molecular evolution. The average of the three 20×20 transition matrices is given in Appendix 1, where amino acids are arranged

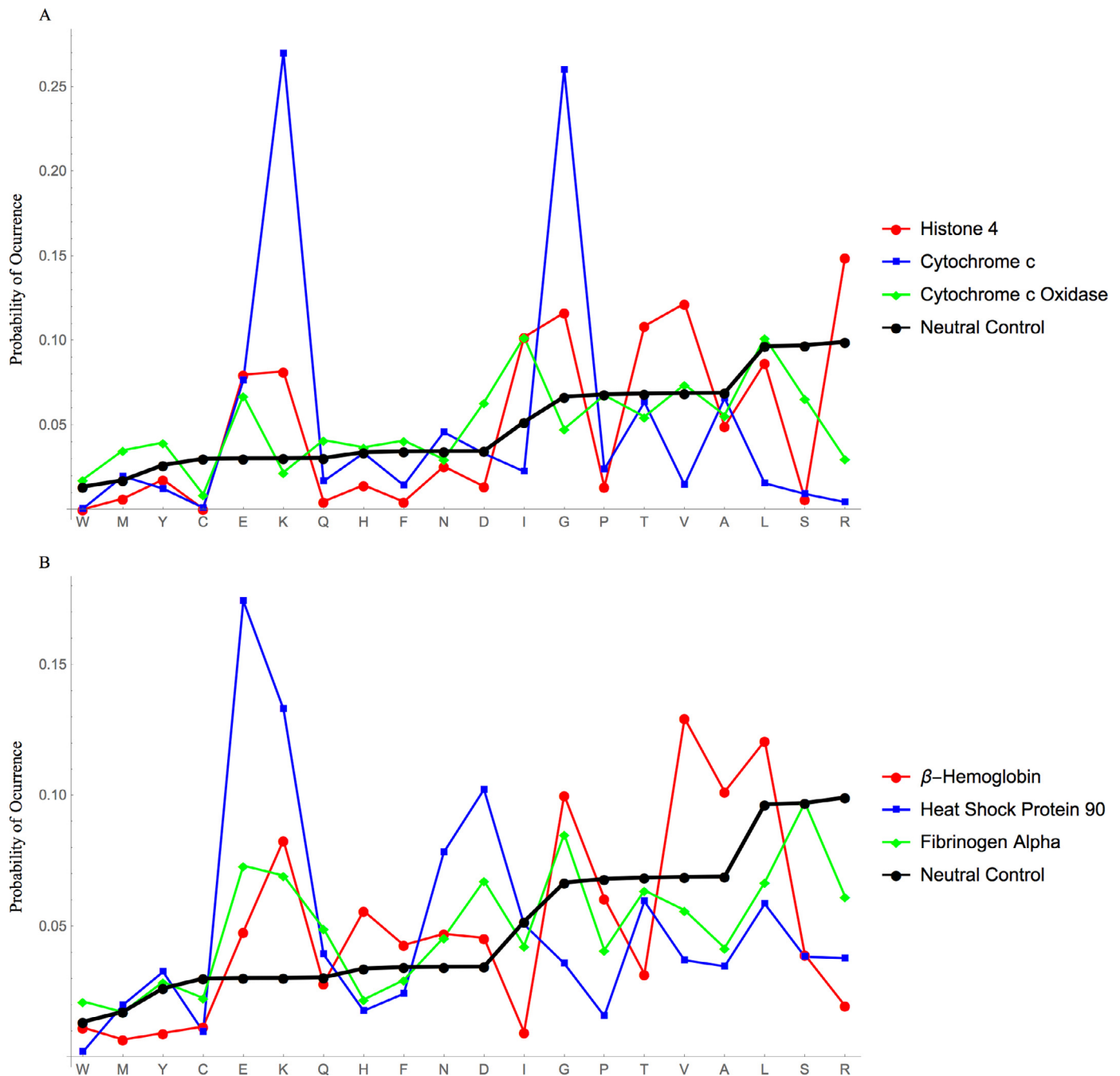


Fig. 2. The amino acid neutrality test (ANT) applied to six proteins: In A), Histone 4 (red-solid circles); Cytochrome c (green-squares); Cytochrome oxidase (blue-rhombi); In B), β -hemoglobin (red-solid circles); Heat shock protein (green-squares); Fibrinogen alpha (blue-rhombi); The neutral control (black-solid circles). The amino acids are ordered according to its degeneracy, i.e., from mono-codonic: W (Trp), M (Met), Y (Tyr); di-codonic: C (Cys), E (Glu), K (Lys), Q (Gln), H (Hist), F (Phe), N (Asn), D (Asp); three-codonic: I (Ile); tetra-codonic: G (Gly), P (Pro), T (Thr), V (Val); A (Ala); and hexacodonic: L (Leu), S (Ser), and R (Arg). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

of sequences used, and the ones derived from the JC-model and BLOSUM62 substitution matrix are presented in Table 1. For the proteins analyzed, the amino acid stationary distribution values from which its corresponding confidence interval is below the neutral control are marked in red, the ones undergoing positive selection are marked in green, and the ones showing neutral selection are in black. In the case of **cytochrome c**, Met, Tyr, Gln, His, Asp, Ile, Thr, Ala are neutral. Glu, Lys, Asn, Gly have positive selection. Trp, Cys, Phe, Pro, Val, Leu, Ser, and Arg show negative

selection. The highly conserved ribonucleoprotein **histone 4** has Trp, Met, Cys, Gln, His, Phe, Asn, Asp, Pro, Ala, and Ser that have negative selection. Leu and Tyr are neutral. Glu, Lys, Ile, Gly, Thr, Val, and Arg show positive selection. In **cytochrome c oxidase**, His, Pro are neutral. Trp, Met, Tyr, Glu, Gln, Phe, Asp, Ile, Val, Leu show positive selection. Cys, Lys, Asn, Gly, Thr, Ala, Ser, and Arg have negative selection. In **β -hemoglobin**, Trp, Tyr, Gln are neutral. Glu, Lys, His, Phe, Asn, Asp, Gly, Val, Ala, Leu show positive selection. Met, Cys, Ile, Pro, Thr, Ser, and Arg have negative

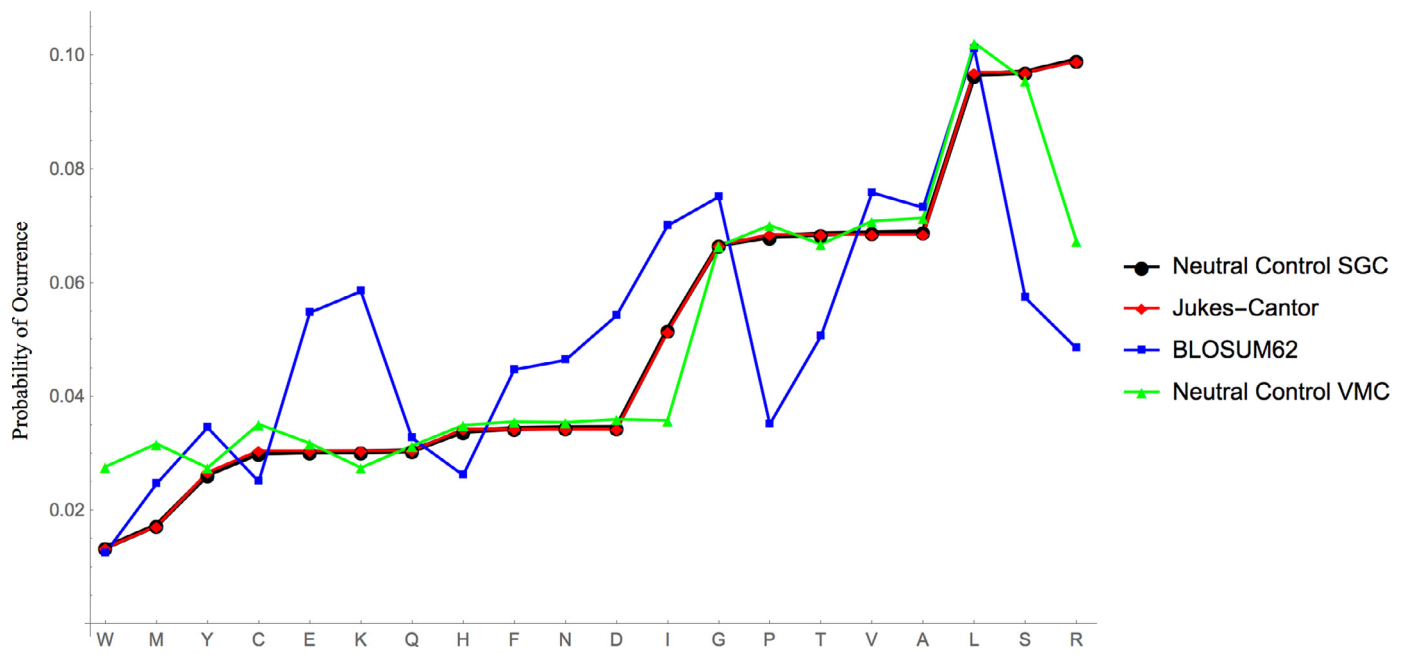


Fig. 3. Stationary probability distributions for the JC-model (diamonds-red curve), BLOSUM62 (squares-blue curve), the neutral control of the SGC (circles-black curve), and the neutral control of the VMC (triangles-green curve). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

selection. In **heat shock protein 90**, Tyr, Gln, Phe, Asn, and Ile are neutral. Met, Glu, Lys, Asp show positive selection. Trp, Cys, His, Gly, Pro, Thr, Val, Ala, Leu, Ser, Arg have negative selection. In **fibrinogen alpha**, Trp, Met, and Ser are neutral. Tyr, Glu, Gln, Lys, Asn, Asp and Gly show positive selection. Cys, His, Phe, Ile, Pro, Thr, Val, Ala, Leu, and Arg have negative selection.

The neutral control is also plotted with the limiting stationary distribution from the model derived from the JC-type substitutions and the BLOSUM62 stochastic process (Fig. 3). The distribution derived from the JC-model is practically indistinguishable from the neutral control. The BLOSUM62 distribution, in contrast, does not maintain the pattern related to the codonicity of each amino acid.

The same procedure used to derive the neutral control for the SGC is used to derive the neutral control for the vertebrate mitochondrial code (VMC). The VMC was downloaded from: <https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?chapter=tgencodes#SG24> (accessed on January 10 2019). The SGC and the VMC have small differences in the assignment of amino acids, only 4 codons are assigned to different amino acids or stop signal. The codon UGA is a stop codon in the SGC whereas in the VMC encodes for Trp; the codon AUA changes from the tri-codonic Ile in the SGC to form a di-codonic Met in the VMC. The codons AGA and AGG change from Arg in the SGC to stop signals in the VMC. These differences in the genetic codes yield to neutral controls with some variations in the probability of occurrence on some amino acids. Particularly the amino acids Cys, Lys, Pro, Thr, Val, Ala, and Leu have small variations; and the amino acids with large variations are Trp, Met, Ile, and Arg. The amino acids with large variations in the probability of occurrence between the two neutral controls for the individual codes are the specific amino acids that have codons with different encodings in the two codes.

5. Discussion

In this work, we introduce a novel amino acid substitution model based on a mathematical model of the SGC. From this substitution model, a test for neutral evolution is derived. This test measures unambiguously the levels of positive or negatively

selected amino acids as well as those that are neutral or close to neutrality. We have ushered in a universal neutral control that depends on the genetic code, from which selective pressures, positive or negative, of a given amino acid can be made, without the need of a phylogenetic tree (Baker et al., 2016). Our results cannot be directly visualized or measured from a phylogenetic tree.

The ANT at amino acid level can be used for assessing the expected amino acid substitution for a single set of protein sequences, as it is compared to the neutral control. We chose widely studied proteins, like β -hemoglobin, or proteins that have slow evolutionary rates like histone 4, in contrast to fibrinogen alpha that has a faster evolutionary rate. In the case of hemoglobin, it has been shown that the rate of amino acid usage in the surface is higher than in the heme pocket (Kimura and Ohta, 1973). It has been shown that there is a negative correlation between the proportion of Gly in a protein and its rate of amino acid usage expectation regardless of its function (Graur, 1985).

Studies of sequence databases have shown that simple genome repeats are abundant in eukaryotic genomes and are considered one of the major sources of genetic variation (Kashi et al., 1997). Single amino acid repeats are not rare in proteins, particularly, glutamine account for a large proportion of them (Green and Wang, 1994). Repeat patterns in protein sequences have been found, although their exact role in protein structure and function has not been fully discerned (Katti et al., 2000). Experiments have shown that Leu is the strongest structure forming and neighboring active-site residue with the possible exception of Met (Chou and Fasman, 1973). Mutations from Leu to a Met residue in proteins may be considered as conservative as both residues are hydrophobic (Némethy and Scheraga, 1962), and are usually placed in highly structured interior regions of proteins near an active site (Chou and Fasman, 1973).

The histone family is the basic set of proteins that coordinate the organization of the eukaryotic DNA into a hierarchical structure known as chromatin (Couture and Trievel, 2006). The nucleosome is the fundamental unit of this structure and its sequence comprises 147 DNA base pairs wound around a histone octamer composed of two copies of the core histones H1, H2A, H2B, H3

and H4 (Luger and Hansen, 2005). The post-translational modifications of the core histones have long been reported (Allfrey et al., 1964). These modifications are an important mechanism for the regulation of the chromatin structure (Goll and Bestor, 2002). Most post-translational modifications reported in histones are found in the NH₂-terminal tail domains (Zhang et al., 2003). These observations have led to the hypothesis of a “histone code” that regulates chromatin structure and accessibility (Strahl and Allis, 2000). Amino acid frequencies have also been reported to impact the functionality in the structure of the histone. It has been shown that the substitution of the Arg 45 residue of histone H4 to cysteine or histidine disrupts the histone-DNA interactions, although, they are more resistant to UV-induced cell death (Nag et al., 2008). Point mutations in histone H4 also have been shown to change the expression of certain locus, as some mutations can mimic the post-translational modifications of the wild type (Park and Szostak, 1990). The specific residue Lys 16 of the N-terminal lysines has been proved to produce a different transcriptional phenotype when mutated (Dion et al., 2005).

The cytochrome c oxidase is a terminal enzyme in the mitochondrial and many bacterial respiratory chains (Lappalainen et al., 1995). It catalyzes the electron transfer between cytochrome c and molecular oxygen (Wikstrom, 1977). Cytochrome c is subdivided into eight or nine subunits (Capaldi, 1982). The residues Asp-112 and Glu-114, in the subunit II, conform a conserved sequence in most species (Capaldi et al., 1983). Other conserved residues are Asp-158, Glu-198, Asp-11, Glu-62, Asp-88, Glu-109, Glu-137, Asp-139 and Asp-173 (Capaldi et al., 1983).

Conserved amino acids of the mitochondrial cytochrome c are of special interest due to their relationship with the prosthetic group (Luntz et al., 1989). The conserved residues Tyr-67, Asn-52, and Thr-78 are considered to be involved in the chemical regulation of the heme group (Takano and Dickerson, 1981a, 1981b). Experiments have shown that these three residues provide the hydrogen bonds to hold an internal molecule of water (Luntz et al., 1989). The relevance of Lys residues in the binding domain of cytochrome c, is that they decrease its stability when these residues are altered (Dopner et al., 1999).

Fibrinogen is an essential protein for platelet aggregation and the formation of fibrin clots; it is made from 3 pairs of polypeptide chains (Collet et al., 2005). Experiments in which the α chains were truncated at residue 251 showed that the α C domains of fibrinogen promote clot stabilization, and have a more prominent role in the mechanical behavior of the fibrin network than in its morphologic properties (Collet et al., 2005). Other experiments have shown that peptides that contain the sequence RGD at positions α 95–97 and α 572–574 inhibit the interaction of fibrinogen with platelet glycoproteins (Farrell et al., 1992).

All organisms respond to heat shifts by synthesizing a group of proteins named the heat-shock proteins (hsp) (Lindquist and Craig, 1988). These proteins act as chaperones and maintain the state and folding of proteins under physiological stress (Morimoto et al., 1990). Most organisms produce proteins from the hsp70 and hsp90 gene families in response to high temperatures, which are among the most conserved proteins in existence (Lindquist and Craig, 1988). The proteins from the hsp90 gene family from eukaryotic organisms share at least a 40% identity with *E. coli* (Bardwell and Craig, 1987). All the eukaryotic proteins from the hsp90 gene family share a conserved sequence of four amino acids Glu-Glu-Val-Asp at the end of the carboxy-terminal regions (Lindquist and Craig, 1988).

It has been shown that hemoglobin is mostly insensitive to mutations in its surface, but it is highly sensitive to alterations in the internal non-polar residues, especially those in contact with the heme group (Perutz and Lehmann, 1968). The role of His-146 in the β -subunit has been widely studied for its relation with

the alkaline Bohr effect, which corresponds to the ability of the hemoglobin molecule to release protons during the transition from deoxy to oxy form (Kwiatkowski and Noble, 1982; Russu et al., 1980). The difference between the distribution derived from the JC-model and the neutral control, reflects the slight tendency of the neutral towards transitions. The contrast between the limiting stationary distribution from BLOSUM62 and the neutral control reflects the effects of natural selection; notice that some amino acids have a higher propensity to stay aligned whereas others are more flexible during the alignment process using the BLOSUM62 matrix.

The 6D representation of the SGC model is derived from a primeval RNY (purine-any-pyrimidine) subcode by means of symmetry breakings (José et al., 2014, 2012, 2007). This model has been proved to be equivalent to the Rodin-Rodin model (Rodin et al., 2011), and the 6D representation displays the symmetrical properties (José et al., 2017) of the distribution of aminoacyl-tRNA synthetases on the amino acids (Rodin et al., 2011) and of polar requirements values (Woese et al., 1966). It has also been used to show the properties of the SGC that, coupled with its evolution from the primeval RNY subcode, determines the uniqueness of the SGC (Zamudio and José, 2017). The use of the limiting distribution of the probability transition matrix as a neutral test is a novel approach to determine the amount of neutral substitutions that occur in a protein sequence as well as positive or negative deviations from neutrality. The removal of the stop codons of the SGC to calculate the transition probabilities of the amino acids has been used in other models (Goldman and Yang, 1994) as mutations of this kind are unlikely to survive.

Amino acid substitution models may be divided according to their focus on nucleotides, codons or amino acids. Theoretically simple models include the JC-model (Jukes and Cantor, 1969) and the Kimura 2-parameter model (Kimura, 1980). These models are built for single nucleotides and do not consider codons or amino acids or the degeneracy of the genetic code. Even more complex and sophisticated theoretical models such as F81 (Felsenstein, 1981), HKY85 (Hasegawa et al., 1985), TN93 (Tamura and Nei, 1993), and the most general model, the generalized time reversal model (Tavaré, 1986) fail to consider the degeneracy of the genetic code, and the distribution of amino acids in the codon table. These parameterized models based on nucleotides developed a mathematical framework to mimic the evolutionary process, rather than setting a control to measure it to a neutral baseline. Codon substitution models as GY94 (Goldman and Yang, 1994) determines the codon-codon substitution rates using a reverse engineering process. GY94 model uses a physicochemical amino acid distance matrix and modifies it to determine codon substitution rates. This model resembles the construction of the 6D representation of the genetic code as the transition rates between two codons is zero if the codons differ by more than one nucleotide; a scaling factor is used to adjust the rates for transitions and transversion which could be directly applied to this model in order to resemble a neutral model obeying the chemical restrictions for transitions and transversions. Other models, such as the MG94 (Muse and Gaut, 1994) uses a codon substitution system that also resembles the 6D representation of the genetic code. This maximum likelihood model determines the codon substitution rates by the factors of synonymous and non-synonymous substitutions coupled with the nucleotide equilibrium frequencies. Both the GY94 and the MG94 models are non-neutral continuous stochastic processes and have properties that consider the degeneracy of the genetic code and its degeneracy in different manners. To our knowledge, no theoretical and mathematical evolutionary model has been developed that explicitly considers codons, amino acids, the degeneracy of the genetic code, and the arrangement of amino acids in the genetic code table as the one presented here. Our

model may be conceived as an evolutionary equivalent of the Hardy-Weinberg equilibrium principle of allelic proportions from population genetics.

The present theoretical model is developed from single nucleotides up to amino acids considering a wide number of properties of the genetic code extending the foundations of the neutral theory of evolution, which is considered as the null hypothesis of molecular evolution (Duret, 2008). So far, Kimura's neutral theory is mathematically modeled for single nucleotides (Kimura, 1981). This substitution model can be useful to correct the mutation rate of proteins and fine-tuning its molecular clock.

Authors contributions

GSZ and MVJ conceived and design the whole work; GSZ derived the theoretical model and performed the experiments and analysis on protein sequences; FC and STF discussed the experimental results and controls; GSZ, FC, STF and MVJ analysed the data, GSZ and MVJ wrote the paper.

Funding

Gabriel S. Zamudio is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and a fellowship recipient from Consejo Nacional de Ciencia y Tecnología (CONACYT) (number: 737920). Marco V. José was financially supported by PAPIIT-IN201019, UNAM, México.

Acknowledgments

We thank Juan R. Bobadilla for technical computer support.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jtbi.2019.01.027.

References

- Allfrey, V.G., Faulkner, R., Mirsky, A.E., 1964. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proc. Natl. Acad. Sci.* 51, 786–794. <https://doi.org/10.1073/pnas.51.5.786>.
- Arenas, M., 2015. Trends in substitution models of molecular evolution. *Front. Genet.* 6, 319. <https://doi.org/10.3389/fgene.2015.00319>.
- Baker, J., Meade, A., Pagel, M., Venditti, C., 2016. Positive phenotypic selection inferred from phylogenies. *Biol. J. Linn. Soc.* 118, 95–115. <https://doi.org/10.1111/bij.12649>.
- Bardwell, J.C., Craig, E.A., 1987. Eukaryotic Mr 83,000 heat shock protein has a homologue in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* 84, 5177–5181. <https://doi.org/10.1073/pnas.84.15.5177>.
- Bressloff, P.C., 2008. Stochastic processes in cell biology.
- Capaldi, R.A., 1982. Arrangement of proteins in the mitochondrial inner membrane. *Biochim. Biophys. Acta - Rev. Biomembr.* 694, 291–306. [https://doi.org/10.1016/0304-4157\(82\)90009-0](https://doi.org/10.1016/0304-4157(82)90009-0).
- Capaldi, R.A., Malatesta, F., Darley-Usmar, V.M., 1983. Structure of cytochrome c oxidase. *Biochim. Biophys. Acta - Rev. Bioenerg.* 726, 135–148. [https://doi.org/10.1016/0304-4173\(83\)90003-4](https://doi.org/10.1016/0304-4173(83)90003-4).
- Chou, P.Y., Fasman, G.D., 1973. Structural and functional role of leucine residues in proteins. *J. Mol. Biol.* 74, 263–281. [https://doi.org/10.1016/0022-2836\(73\)90372-0](https://doi.org/10.1016/0022-2836(73)90372-0).
- Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., De Hoon, M.J.L., 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. <https://doi.org/10.1093/bioinformatics/btp163>.
- Collet, J.P., Moen, J.L., Veklich, Y.I., Gorkun, O.V., Lord, S.T., Montalescot, G., Weisel, J.W., 2005. The α C domains of fibrinogen affect the structure of the fibrin clot, its physical properties, and its susceptibility to fibrinolysis. *Blood* 106, 3824–3830. <https://doi.org/10.1182/blood-2005-05-2150>.
- Couture, J.-F., Trievel, R.C., 2006. Histone-modifying enzymes: encrypting an enigmatic epigenetic code. *Curr. Opin. Struct. Biol.* 16, 753–760. <https://doi.org/10.1016/j.sbi.2006.10.002>.
- Dayhoff, M.O., Schwartz, R., Orcutt, B.C., 1978. A model of evolutionary change in proteins. *Atlas Protein Seq. Struct.* 345–352.
- Dion, M.F., Altschuler, S.J., Wu, L.F., Rando, O.J., 2005. Genomic characterization reveals a simple histone H4 acetylation code. *Proc. Natl. Acad. Sci.* 102, 5501–5506. <https://doi.org/10.1073/pnas.0500136102>.
- Dopner, S., Hildebrandt, P., Rosell, F.I., Mauk, A.G., von Walter, M., Buse, G., Soulimane, T., 1999. The structural and functional role of lysine residues in the binding domain of cytochrome c in the electron transfer to cytochrome c oxidase. *Eur. J. Biochem.* 261, 379–391. <https://doi.org/10.1046/j.1432-1327.1999.00249.x>.
- Duret, L., 2008. Neutral theory: the null hypothesis of molecular evolution. *Nat. Educ.* 1, 1–5.
- Eddy, S.R., 2004. Where did the BLOSUM62 alignment score matrix come from? *Nat. Biotechnol.* <https://doi.org/10.1038/nbt0804-1035>.
- Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. <https://doi.org/10.1093/nar/gkh340>.
- Farrell, D.H., Thiagarajan, P., Chung, D.W., Davie, E.W., 1992. Role of fibrinogen alpha and gamma chain sites in platelet aggregation. *Proc. Natl. Acad. Sci. USA* 89, 10729–10732. <https://doi.org/10.1073/pnas.89.22.10729>.
- Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376. <https://doi.org/10.1007/BF01734359>.
- Goldman, N., Yang, Z., 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11, 725–736. <https://doi.org/10.1093/oxfordjournals.molbev.a040153>.
- Goll, M.G., Bestor, T.H., 2002. Histone modification and replacement in chromatin activation. *Genes Dev.* <https://doi.org/10.1101/gad.1013902>.
- Graur, D., 1985. Amino acid composition and the evolutionary rates of protein-coding genes. *J. Mol. Evol.* 22, 53–62. <https://doi.org/10.1007/BF02105805>.
- Green, H., Wang, N., 1994. Codon reiteration and the evolution of proteins. *Proc. Natl. Acad. Sci. USA* 91, 4298–4302. <https://doi.org/10.1073/pnas.91.10.4298>.
- Hasegawa, M., Kishino, H., Yano, T., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22, 160–174. <https://doi.org/10.1007/BF02101694>.
- Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* 89, 10915–10919. <https://doi.org/10.1073/pnas.89.22.10915>.
- José, M.V., Morgado, E.R., Govezensky, T., 2007. An extended RNA code and its relationship to the standard genetic code: an algebraic and geometrical approach. *Bull. Math. Biol.* 69, 215–243. <https://doi.org/10.1007/s11538-006-9119-3>.
- José, M.V., Morgado, E.R., Guimarães, R.C., Zamudio, G.S., de Farias, S.T., Bobadilla, J.R., Sosa, D., 2014. Three-dimensional algebraic models of the tRNA code and 12 graphs for representing the amino acids. *Life (Basel, Switzerland)* 4, 341–373. <https://doi.org/10.3390/life4030341>.
- José, M.V., Morgado, E.R., Sanchez, R., Govezensky, T., 2012. The 24 possible algebraic representations of the standard genetic code in six or in three dimensions. *Adv. Stud. Biol.* 4, 119–152.
- José, M.V., Zamudio, G.S., Morgado, E.R., 2017. A unified model of the standard genetic code. *R. Soc. Open Sci.* 4, 160908. <https://doi.org/10.1098/rsos.160908>.
- José, M.V., Zamudio, G.S., Palacios-Pérez, M., Bobadilla, J.R., de Farias, S.T., 2015. Symmetrical and thermodynamic properties of phenotypic graphs of amino acids encoded by the primeval RNY code. *Orig. Life Evol. Biosph.* 45, 77–83. <https://doi.org/10.1007/s11084-015-9427-4>.
- Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules, in: *mammalian protein metabolism*.
- Kashi, Y., King, D., Soller, M., 1997. Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet.* 13, 74–78. [https://doi.org/10.1016/S0168-9525\(97\)01008-1](https://doi.org/10.1016/S0168-9525(97)01008-1).
- Katti, M.V., Sami-Subbu, R., Ranjekar, P.K., Gupta, V.S., 2000. Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications. *Protein Sci.* 9, 1203–1209. <https://doi.org/10.1110/ps.9.6.1203>.
- Kimura, M., 1991. Recent development of the neutral theory viewed from the Wrightian tradition of theoretical population genetics. *Proc. Natl. Acad. Sci.* 88, 5969–5973. <https://doi.org/10.1073/pnas.88.14.5969>.
- Kimura, M., 1981. Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA* 78, 454–458. <https://doi.org/10.1073/pnas.78.1.454>.
- Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120. <https://doi.org/10.1007/BF01731581>.
- Kimura, M., 1979. The neutral theory of molecular evolution. 241, 94–104.
- Kimura, M., Ohta, T., 1973. Mutation and evolution at the molecular level. *Genetics*.
- Kwiatkowski, L.D., Noble, R.W., 1982. The contribution of histidine (HC3) (146 beta) to the R state Bohr effect of human hemoglobin. *J. Biol. Chem.* 257, 8891–8895.
- Lappalainen, P., Saraste, M., Watmough, N.J., Greenwood, C., 1995. Electron transfer between cytochrome c and the isolated CuA domain: identification of substrate-binding residues in cytochrome c oxidase. *Biochemistry* 34, 5824–5830. <https://doi.org/10.1021/bi00017a014>.
- Lindquist, S., Craig, E.A., 1988. The heat-shock proteins. *Annu. Rev. Genet.* 22, 631–677. <https://doi.org/10.1146/annurev.ge.22.120188.003215>.
- Luger, K., Hansen, J.C., 2005. Nucleosome and chromatin fiber dynamics. *Curr. Opin. Struct. Biol.* <https://doi.org/10.1016/j.sbi.2005.03.006>.
- Luntz, T.L., Schejter, A., Garber, E.a, Margoliash, E., 1989. Structural significance of an internal water molecule studied by site-directed mutagenesis of tyrosine-67 in rat cytochrome c. *Proc. Natl. Acad. Sci. USA* 86, 3524–3528. <https://doi.org/10.1073/pnas.86.10.3524>.

- Morimoto, R.I., Tissières, A., Georgopoulos, C., 1990. The stress response, function of the proteins, and perspectives. *Cold Spring Harb. Monogr. Arch.* 19, 1–36. <https://doi.org/10.1101/087969337.19.1>.
- Muse, S.V., Gaut, B.S., 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* 11, 715–724. <https://doi.org/10.1093/oxfordjournals.molbev.a040152>.
- Nag, R., Gong, F., Fahy, D., Smerdon, M.J., 2008. A single amino acid change in histone H4 enhances UV survival and DNA repair in yeast. *Nucleic Acids Res.* 36, 3857–3866. <https://doi.org/10.1093/nar/gkn311>.
- Némethy, G., Scheraga, H.A., 1962. Structure of water and hydrophobic bonding in proteins. II. Model for the thermodynamic properties of aqueous solutions of hydrocarbons. *J. Chem. Phys.* 36, 3401–3417. <https://doi.org/10.1063/1.1732473>.
- Ng, P.C., Henikoff, S., 2006. Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.* 7, 61–80. <https://doi.org/10.1146/annurev.genom.7.080505.115630>.
- Ohta, T., 1973. Slightly deleterious mutant substitutions in evolution. *Nature* 246, 96–98. <https://doi.org/10.1038/246096a0>.
- Pál, C., Papp, B., Lercher, M.J., 2006. An integrated view of protein evolution. *Nat. Rev. Genet.* 7, 337–348. <https://doi.org/10.1038/nrg1838>.
- Park, E.C., Szostak, J.W., 1990. Point mutations in the yeast histone H4 gene prevent silencing of the silent mating type locus HML. *Mol. Cell. Biol.* 10, 4932–4934. <https://doi.org/10.1128/MCB.10.9.4932>.
- Perutz, M.F., Lehmann, H., 1968. Molecular pathology of human haemoglobin. *Nature* 219, 902–909. <https://doi.org/10.1038/219902a0>.
- Rodin, A.S., Szathmáry, E., Rodin, S.N., 2011. On origin of genetic code and tRNA before translation. *Biol. Direct* 6, 14. <https://doi.org/10.1186/1745-6150-6-14>.
- Russu, I.M., Ho, N.T., Ho, C., 1980. Role of the .beta.146 histidyl residue in the alkaline Bohr effect of hemoglobin. *Biochemistry* 19, 1043–1052. <https://doi.org/10.1021/bi00546a033>.
- Sigman, K., 1995. *Stationary Marked Point Processes, an Intuitive Approach*. Chapman & Hall.
- Strahl, B.D., Allis, C.D., 2000. The language of covalent histone modifications. *Nature*. <https://doi.org/10.1038/47412>.
- Takano, T., Dickerson, R.E., 1981a. Conformation change of cytochrome c. I. Ferrocyanide structure refined at 1.5 Å resolution. *J. Mol. Biol.* 153, 79–94. [https://doi.org/10.1016/0022-2836\(81\)90528-3](https://doi.org/10.1016/0022-2836(81)90528-3).
- Takano, T., Dickerson, R.E., 1981b. Conformation change of cytochrome c. II. Ferricytochrome c refinement at 1.8 Å and comparison with the ferrocyanide structure. *J. Mol. Biol.* 153, 95–115. [https://doi.org/10.1016/0022-2836\(81\)90529-5](https://doi.org/10.1016/0022-2836(81)90529-5).
- Tamura, K., Nei, M., 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10, 512–526. <https://doi.org/10.1093/oxfordjournals.molbev.a040023>.
- Tavaré, S., 1986. In: *Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences*, in: *American Mathematical Society: Lectures on Mathematics in the Life Sciences, c1986*. Providence, R.I. American Mathematical Society, pp. 57–86.
- The UniProt Consortium, 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 45, D158–D169. <https://doi.org/10.1093/nar/gkw1099>.
- Wikstrom, M.K.F., 1977. Proton pump coupled to cytochrome c oxidase in mitochondria. *Nature*. <https://doi.org/10.1038/266271a0>.
- Woese, C.R., Dugre, D.H., Saxinger, W.C., Dugre, S.A., 1966. The molecular basis for the genetic code. *Proc. Natl. Acad. Sci. USA* 55, 966–974. <https://doi.org/10.1073/pnas.55.4.966>.
- Zamudio, G.S., José, M.V., 2017. On the uniqueness of the standard genetic code. *Life* 7, 7. <https://doi.org/10.3390/life7010007>.
- Zhang, L., Eugeni, E.E., Parthun, M.R., Freitas, M.A., 2003. Identification of novel histone post-translational modifications by peptide mass fingerprinting. *Chromosoma* 112, 77–86. <https://doi.org/10.1007/s00412-003-0244-6>.

Apéndice 8

Article

On the Importance of Asymmetry in the Phenotypic Expression of the Genetic Code upon the Molecular Evolution of Proteins

Marco V. José *  and Gabriel S. Zamudio 

Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, Ciudad Universitaria 04510, Mexico; gazaso_92@comunidad.unam.mx

* Correspondence: marcojose@biomedicas.unam.mx

Received: 28 April 2020; Accepted: 20 May 2020; Published: 11 June 2020



Abstract: The standard genetic code (SGC) is a mapping between the 64 possible arrangements of the four RNA nucleotides (C, A, U, G) into triplets or codons, where 61 codons are assigned to a specific amino acid and the other three are stop codons for terminating protein synthesis. Aminoacyl-tRNA synthetases (aaRSs) are responsible for implementing the SGC by specifically amino-acylating only its cognate transfer RNA (tRNA), thereby linking an amino acid with its corresponding anticodon triplets. tRNAs molecules bind each codon with its anticodon. To understand the meaning of symmetrical/asymmetrical properties of the SGC, we designed synthetic genetic codes with known symmetries and with the same degeneracy of the SGC. We determined their impact on the substitution rates for each amino acid under a neutral model of protein evolution. We prove that the phenotypic graphs of the SGC for codons and anticodons for all the possible arrangements of nucleotides are asymmetric and the amino acids do not form orbits. In the symmetrical synthetic codes, the amino acids are grouped according to their codonicity, this is the number of triplets encoding a given amino acid. Both the SGC and symmetrical synthetic codes exhibit a probability of occurrence of the amino acids proportional to their degeneracy. Unlike the SGC, the synthetic codes display a constant probability of occurrence of the amino acid according to their codonicity. The asymmetry of the phenotypic graphs of codons and anticodons of the SGC, has important implications on the evolutionary processes of proteins.

Keywords: standard genetic code; symmetry; asymmetry; anticodon code; phenotypic graphs; protein evolution

1. Introduction

The decipherment of the standard genetic code (SGC) is a landmark achievement in biological sciences [1,2]. The SGC is a nearly universal map of 61 nucleotide triplets (codons) to 20 amino acids plus two punctuation marks (three stop and one start signals). The SGC became an abstract mathematical problem even before its discovery [3,4]. Symmetrical properties of the SGC have been found [5–8]. Protein synthesis is the outcome of a complex translation system that involves ribozymes, ribosomal proteins, aminoacyl-tRNA synthetases (aaRSs), elongation and termination factors, and three kinds of RNA molecules, to wit, messenger RNA (mRNA), ribosomal RNA (rRNA), and transfer RNA (tRNA) [9,10]. The evolution of tRNAs, aaRSs, and the SGC has been thoroughly examined and reviewed elsewhere [11,12]. Different models for the evolution of the genetic code have been proposed based on different properties of the amino acids and the triplets [13]. Some models consider an RNY primeval code from which the SGC can be derived by frameshift reading mistranslations and/or transitions and transversions in the first or third nucleotide of each codon [7,14,15]. Other models consider

the metabolic pathways for the incorporation of amino acids to the code [16]. Other propositions are the co-evolution model theory [17–20], and Trifonov’s consensus model [21].

The 20 encoded amino acids exhibit unique physicochemical properties, which facilitate folding, catalysis, and solubility of proteins, and confer adaptive value to organisms able to encode them [22]. Experimental scientists are generally not attracted/interested in theoretical works, which are rarely cited by biologists as they are mathematically abstract and often divorced from biological context [23]. Despite connections between the mathematical models and data [24], theoretical approaches are still regarded as speculative work [25]. It has been suggested that part of the problem lies in the fact that theorists focused on the table of the SGC [26] and failed to address the central question: co-ordinate evolution of aaRS gene sequences and their cognate tRNAs with the codon assignments [25]. To our knowledge, there are theoretical works that tackle the anticodon and operational codes that are found embedded in tRNA molecules [27–29].

In previous works [7,30], Genetic Hotels of codons and Hotels of amino acids (three-dimensional models) were used to test hypotheses about the evolution of the SGC [14,24]. The usual representation of the SGC as a table allows the visualization of the wobble effect that confers robustness to the genetic code by relating similar codons to the same amino acid, most commonly allowing variation in the third position of the codon triplets and fixing the other two bases, with the exception of the hexa-codonic amino acids. The mathematical representation of this effect has been described previously [15] with the computation of the group of automorphisms of the 6D model of the codons that maintain invariant all the equivalent classes of the codons given by the genetic code. It was shown that these groups are not trivial, thus, providing a theoretical representation of the wobbling effect. In this work, we are concerned with the organization of the amino acids in the SGC.

In the present work, we compare the SGC with the standard tRNA code (S-tRNA-C), which comprises 45 tRNAs (an A in the first anticodon position does not exist and there are no anticodons for stop codons). Codon-anticodon pairing takes place according to wobbling rules in which anticodons recognize more than one codon [26]. Hence, the number of required anticodons is reduced substantially. Overall, anticodons beginning with purines (R) are always of only one kind for one amino acid; this is usually G, sometimes a modified purine. This remarkable wobbling property permits that two or more neighboring codon triplets share a common anticodon. Degeneracy is the known property of SGC of having different numbers of codons specifying each amino acid, also named codonicity. For example, Methionine and Tryptophan are specified by a single codon; and Leucine, Arginine, and Serine are encoded by six triplets. The other 15 amino acids have intermediate values with their codonicity ranging from two to four.

In this work, we set out to search for symmetries of the phenotypic graphs of codons and anticodons, and to discern the biological meaning of symmetry. As the phenotypic graphs (PHGs) of anticodons and codons were found to be asymmetric, we designed synthetic and symmetrical codes (sy^2 -codes) with the same degeneracy of the SGC.

We applied a neutral model of evolution to proteins [31] as they would be obtained by the biological anticodon code and the built-in sy^2 -codes. In the sy^2 -codes, subsets of amino acids formed orbits according to their codonicity, whilst in the natural code, 20 orbits were observed, i.e., one orbit for each amino acid.

Definition 1. *The orbit of an element $x \in E$, is defined as $Orb(x) := \{y \in E : \exists g \in G : y = g * x\}$, where $*$ denotes the group action. Hence $Orb(x) = G * x$. The latter means that the orbit of an element is all its possible images or destinations under the group action.*

Definition 2. *Let \mathfrak{R} be the relation on E defined as $\forall x, y \in E : x \mathfrak{R} y \Leftrightarrow \exists g \in G : y = g * x$, where $*$ denotes the group action. The orbit of x , denoted by $Orb(x)$, is the equivalence class of x under \mathfrak{R} .*

The implication of the asymmetry now becomes apparent: the occurrence of each amino acid in protein evolution is independent of the presence/absence of the remaining 19 amino acids. This means

that the process of molecular evolution applied to proteins sequences becomes free and independent from the strict rules dictated by the SGC due to the selected asymmetry—or lack of symmetry—of the graph of codons and anticodons. In other words, the asymmetry of the anticodon code is disassociated with the deterministic character of the SGC.

2. Materials and Methods

The SGC has been modeled upon a 6D hypercube as template using group theory [7,15,32]. The vertices of the hypercube represent the 64 possible nucleotide triplets and the edges join triplets that differ by one nucleotide under different arrangements of the nucleotides in a square (Figure 1). Each of the three possible arrangements of the nucleotides in the square yields different orderings of the codon triplets in the hypercube [33]; a fourth arrangement of the nucleotides is given by the square with its two diagonals, representing a scenario where all possible nucleotide changes are within reach in one step mutation. A more detailed description of the four possible arrangements and their corresponding modes of evolution has been reported elsewhere [31]. All 64 triplets can be represented in a 6D hypercube by considering the four possible arrangements in a square of the four nucleotides [7,15]. The SGC can be readily visualized as a graph of vertices, representing the codons, and edges, joining the codons at the Hamming distance of one. Thereby, the symmetries of the SGC can be obtained from the group of automorphisms of the graph [15].

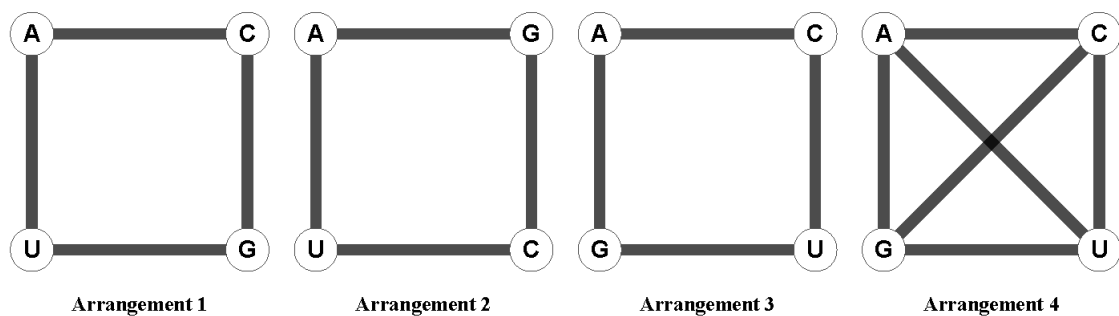


Figure 1. Four possible arrangements of the four nucleotides (A, C, G, U) as the vertices of a square. The four arrangements are not symmetrical but equivalent to the standard genetic code (SGC) and are synthetic symmetrical genetic codes.

The 6D model of the SGC is further transformed into its corresponding amino acid PHG through the algebraic quotient of the 6D model as a graph with the equivalence relation given by the assignation of codons to its corresponding amino acid [27,33,34]. The codons are removed from the hypercube model to correctly produce a PHG where the vertices identify the set of amino acids, and any two given amino acids are joined by an edge if in the 6D hypercube model there are codons differing for such amino acids that were previously joined. The symmetries of a PHG are given by the group of graphs automorphisms which are all the bijective transformations of a graph to itself that preserves adjacencies.

The construction of the codon hypercube model and its corresponding PHG is also computed for the set of anticodon triplets which consist of the set of reverse complementary triplets of the codons with the removal of the anticodon triplets starting with adenine and the anticodons corresponding to the codons for the stop signal, thus resulting in a set of 45 anticodons.

The 6D codon model combined with the PHGs has been used to calculate the probability transition matrix of a stochastic process that models amino acid substitutions given by a neutral model on which the nucleotide changes are at random [31]. We remark that the neutral evolution model and the neutral mutation are as universal as the SGC [31]. The three stochastic processes using the three possible arrangements of the nucleotides in the square with their diagonals are calculated and averaged. The stationary distribution of the averaged stochastic process determines the probability of finding an amino acid, provided that the protein mutates without selection pressures. The averaged stochastic

Second base	U	C	A	G
U	Leu	Ser	Arg	Val
C	Leu	Pro	Arg	Gly
A	Ile	Thr	Asp	Ile
G	Val	Ala	Glu	Met
U	Pro	Thr	Ala	Met
C	Pro	Thr	Ala	Met
A	Pro	Thr	Ala	Met
G	Pro	Thr	Ala	Met

The standard genetic code table a) and the tables of three symmetric synthetic genetic codes. Changes from the synthetic code 1 b) to the synthetic code 2 c) and the synthetic code 3 d) are shaded in grey.

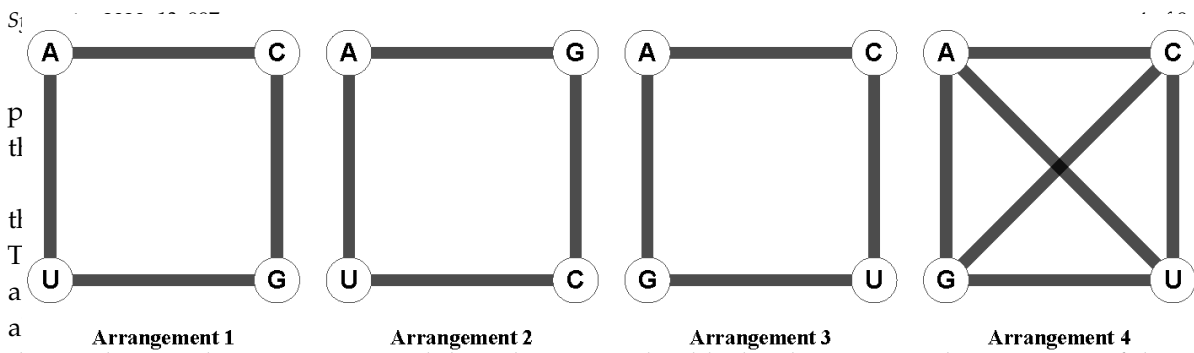


Figure 2. The standard genetic code (SGC) and three synthetic symmetrical genetic codes. The automorphism groups of their PHGs are computed for the four arrangements of the nucleotides. The stochastic process and stationary distributions for each code are further derived.

a) Standard Genetic Code

		Second base				Third base
		U	C	A	G	
First base	U	Phe	Ser	Tyr	Cys	U C A G
	Leu	Ser	Stop	Stop	Irp	
	C	Leu	Pro	His	Arg	
			Gln			
	A	Ile	Thr	Asn	Ser	
	Met		Lys	Arg		
G	Val	Ala	Asp	Gly		
		Glu				

b) Sy²-code 1

		Second base				Third base
		U	C	A	G	
First base	U	Leu	Ser	Arg	Val	U C A G
	Leu	Ser	Arg	Gly		
	C	Phe	Tyr	Cys		
	A	His	Asn	Asp	Ile	
	Gln	Lys	Glu	Met		
	G	Pro	Thr	Ala	Stop	
			Irp			

c) Sy²-code 2

		Second base				Third base
		U	C	A	G	
First base	U	Leu	Ser	Arg	Val	U C A G
	Leu	Ser	Arg	Gly		
	C	Phe	Tyr	Cys		
	A	His	Asn	Asp	Ile	
	Gln	Lys	Glu	Stop		
	G	Pro	Thr	Ala	Stop	
			Met	Irp		

d) Sy²-code 3

		Second base				Third base
		U	C	A	G	
First base	U	Leu	Ser	Arg	Val	U C A G
	Leu	Ser	Arg	Tyr		
	C	Phe	Stop	Cys	Stop	
		Irp				
	A	His	Asn	Asp	Ile	
	Gln	Lys	Glu	Met		
G	Pro	Thr	Ala	Gly		

Figure 2. The standard genetic code (SGC) and the tables of three symmetric synthetic genetic codes. Changes from the standard genetic code table (a) to the synthetic code 1 (b) and the synthetic code 2 (c) and the synthetic code 3 (d) are shaded in grey. The standard genetic code table (a) and the tables of three symmetric synthetic genetic codes. Changes from the synthetic code 1 (b) to the synthetic code 2 (c) and the synthetic code 3 (d) are shaded in grey.

3. Results

The PHGs for the SGC, and the synthetic codes using both codons and anticodons were computed using the four possible arrangements of nucleotides and analyzed for their symmetries (Table 1). The PHGs for the SGC for codons and anticodons for all the arrangements have as symmetry group the trivial group e , which means that there are no symmetric transformations other than the identity transformation. The PHGs for the synthetic code 1 on both codons and anticodons have as symmetry the group \mathbb{Z}_2 for the arrangements of nucleotides without diagonals (arrangements 1, 2, and 3), whereas for the arrangement with diagonals (arrangement 4) the symmetry group is given by S_3 . For the synthetic code 2 the symmetry group of the codons is S_3 and the symmetry group of the anticodons is S_3 when considering the arrangement 2 of the nucleotides, the PHG of the codons has as symmetry the group \mathbb{Z}_2 whereas for the anticodon PHG the symmetry group is \mathbb{Z}_2 . The symmetry groups of the codons and anticodons coincide for the other nucleotide arrangements. For the synthetic code 3 the codon and anticodon PHGs coincide, with only the arrangement 2 without symmetries and the other nucleotide arrangements

having the symmetry group given by \mathbb{Z}_2 . The group given by \mathbb{Z}_2 geometrically represents a reflection through an axis, the group \mathbb{Z}_2^2 represents the symmetries of a rectangle, whereas the group S_3 represents the permutations of a set with three elements or the symmetries of an equilateral triangle.

Table 1. Symmetry groups.

	Standard Genetic Code	Synthetic Code 1	Synthetic Code 2	Synthetic Code 3
	Phenotypic graph of codons/anticodons	Phenotypic graph of codons/anticodons	Phenotypic graph of codons/anticodons	Phenotypic graph of codons/anticodons
Arrangement 1	e	\mathbb{Z}_2	\mathbb{Z}_2^2	\mathbb{Z}_2
Arrangement 2	e	\mathbb{Z}_2	\mathbb{Z}_2^2	e
Arrangement 3	e	\mathbb{Z}_2	\mathbb{Z}_2	\mathbb{Z}_2
Arrangement 4	e	S_3	S_3	\mathbb{Z}_2

The sets of orbits of each PHG under the action of its symmetry group are shown in Table 2. For the PHGs of codons and anticodons of the SGC, there are no orbits in any of the four arrangements. For the synthetic symmetrical codes, note that the codonicity of the amino acids grouped through the action of the symmetry group is the same, i.e., given an amino acid, its orbit under the action of the group contains amino acids which have the same codonicity. The latter does not hold true for the synthetic code 3, where the graphs of codons and anticodons are asymmetric under arrangement 2.

Table 2. Sets of orbits.

	Standard Genetic Code	Synthetic Code 1	Synthetic Code 2	Synthetic Code 3
	Phenotypic graph of codons/anticodons	Phenotypic graph of codons/anticodons	Phenotypic graph of codons/anticodons	Phenotypic graph of codons/anticodons
Arrangement 1	N/A	{N, H}, {Q, K}, {L, S}, {E, Y}, {P, T}	{N, H}, {Q, K}, {L, S}, {E, Y}, {P, T}	{N, H}, {Q, K}, {L, S}, {E, W}, {P, T}
Arrangement 2	N/A	{A, T}, {R, S}, {N, D}, {C, Y}, {E, K}	{A, T}, {R, S}, {N, D, E, K}/A, T, {R, S}, {N, D}, {C, Y}, {Q, H}, {C, Y}, {E, K}	N/A
Arrangement 3	N/A	{A, P}, {R, L}, {D, H}, {C, F}, {Q, E}	{A, P}, {R, L}, {D, H}, {C, F}, {Q, E}	{A, P}, {R, L}, {D, H}, {C, F}, {Q, E}
Arrangement 4	N/A	{A, P, T}, {R, L, S}, {N, D, H}, {C, E, Y}, {Q, E, K}	{A, P, T}, {R, L, S}, {N, D, H}, {C, E, Y}, {Q, E, K}	{A, P}, {R, L}, {D, H}, {C, F}, {Q, E}

Mono-codonic: W (Trp), M (Met); di-codonic: Y (Tyr), C (Cys), E (Glu), K (Lys), Q (Gln), H (His), F (Phe), N (Asn), D (Asp); tri-codonic: I (Ile); tetra-codonic: A (Ala), T (Thr), G (Gly), P (Pro), V (Val); and hexa-codonic: L (Leu), S (Ser), and R (Arg).

The codon and anticodon graphs facilitate the analyses of the evolvability of the genetic code, including tailored-design codes. The stationary distributions of the stochastic processes retrieved by each genetic code are shown in Figure 3. In the stationary distribution for the synthetic code 1, the probability of the amino acids with the same codonicity is approximately the same with a variance of 7.29×10^{-8} for the amino acids with two codons (Tyr (Y), Cys (C), Glu (E), Lys (K), Gln (Q), His (H), Phe (F), Asn (N), Asp (D)); 2.12×10^{-10} for the amino acids with four codons (Gly (G), Pro (P), Thr (T), Val (V), Ala (A)); and 2.76×10^{-7} for the amino acids with six codons (Leu (L), Ser (S), Arg (R)). This pattern is not present for the synthetic codes 2 and 3. For the synthetic code 2 the amino acids as Glu, Lys, Gln, Gly, and Val have lower probabilities than the other amino acids with the same codonicities. For the synthetic code 3 the amino acids Tyr, Cys, Phe, Gly, Val, and Ser have lower probabilities than the other amino acids with the same codonicities. The equiprobable behavior of the stationary distribution of the synthetic code 1 is not present in the stationary distribution of the SGC where the probabilities of Tyr, Cys, Glu, Lys, Gln are lower than the probabilities of His, Phe, Asn, Asp for the di-codonic amino acids. On the amino acids with four codons the probabilities of Gly and Pro are lower than the probabilities of Thr, Val, Ala. For the hexa-codonic amino acids the probability of Arg is greater than the probability of Leu and Ser. On the synthetic codes and the SGC the probabilities for the uni-coded amino acids Trp and Met are significantly different due to the removal of the codons

Arrangement		{A, P, T}, {R, L, S},	{A, P, T}, {R, L, S},	{A, P}, {R, L}, {D, H}, {C, F}, {Q, E}
4	N/A	{N, D, H}, {C, F, Y}, {Q, E, K}	{N, D, H}, {C, F, Y}, {Q, E, K}	

Mono-codonic: W (Trp), M (Met); di-codonic: Y (Tyr), C (Cys), E (Glu), K (Lys), Q (Gln), H (His), F (Phe), N (Asn), D (Asp); tri-codonic: I (Ile); tetra-codonic: A (Ala), T (Thr), G (Gly), P (Pro), V (Val); and hexa-codonic: L (Leu), S (Ser), and R (Arg).

for the stop signal in the construction of the PHGs. Exact values of the stationary distributions are provided in Supplementary Information S10C or the Synthetic Codes.

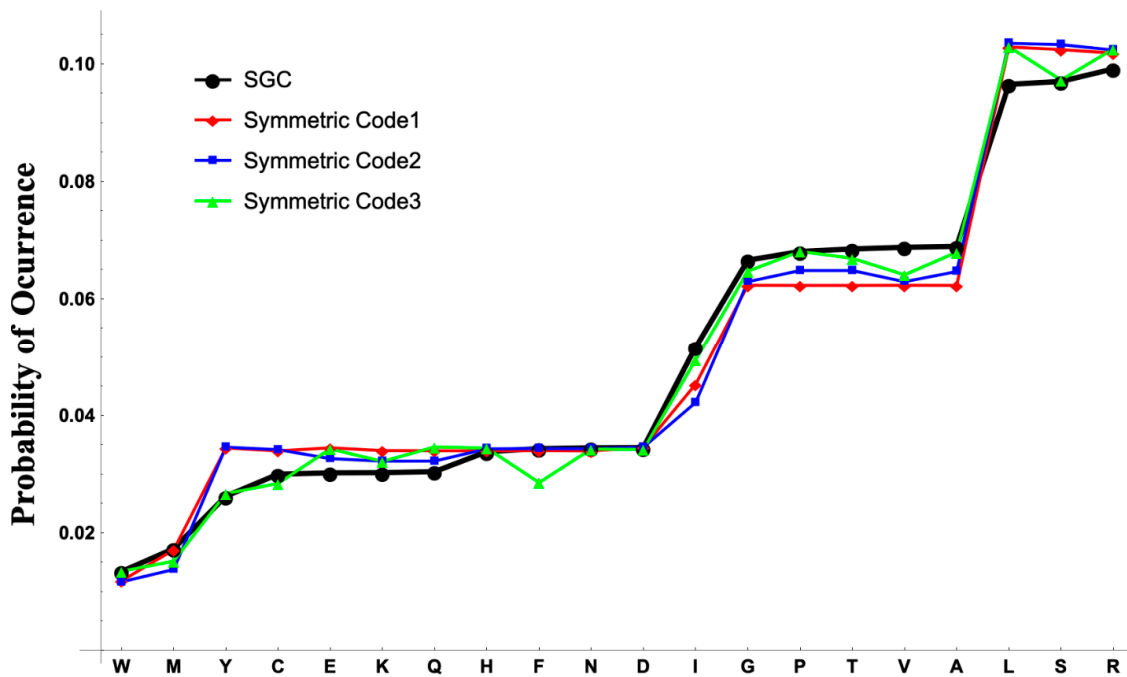


Figure 3. The Amino Acid Neutrality Test Applied to a Hypothetical Protein Obeying the SGC or the Synthetic Codes. The amino acids are ordered according to their codonicity, i.e., from mono-codonic: W (Trp), M (Met); di-codonic: Y (Tyr), C (Cys), E (Glu), K (Lys), Q (Gln), H (His), F (Phe), N (Asn), D (Asp); tri-codonic: I (Ile); tetra-codonic: G (Gly), P (Pro), T (Thr), V (Val); A (Ala); and hexa-codonic: L (Leu), S (Ser), and R (Arg).

4. Discussion

In this work we showed that the PHGs of codons and anticodons as obtained from the SGC are asymmetric for all possible arrangements of the nucleotides. This result is in stark contrast with the symmetries found in the SGC [5–8]. To elucidate the meaning of the asymmetry observed, three synthetic and symmetrical codes with non-trivial symmetries on the mathematical representation of codons and anticodons were designed. We found that the corresponding PHGs of codons and anticodons exhibited symmetry except for arrangement 2. We also observed that in the built-in symmetric codes the amino acids were grouped in similar orbits according to their codonicity, i.e., the amino acids contained in the same orbit have the same codonicity. In contrast, in the natural PHGs the amino acids are orbit-free. Recall that the degeneracy of both the SGC and the synthetic symmetrical codes is the same. Yet, the PHGs of the SGC are asymmetric whereas most PHG of the synthetic symmetrical codes are symmetrical. Next, we subjected the four codes to a neutral model of protein evolution and calculated their stationary distributions. The probability of occurrence of the amino acids for the synthetic codes have a constant probability determined by the redundancy of the amino acid. For the case of the synthetic code 1, whose symmetries are non-trivial for all the nucleotide arrangements and the same codon and anticodon representation in 6D, the amino acids on the same orbit have the same probability of appearance as shown by their respective stationary distribution of the amino acid substitution process. This phenomenon is not as clear for the synthetic code 3, where the symmetry group for the codon and anticodon representation is the trivial group. The neutral test is the null hypothesis of evolution, and it clearly discerns putative amino acids with unexpected frequencies that might be under positive, negative selection, or neutral. Therefore, the $\frac{d_N}{d_S}$ ratio, i.e., the number of nonsynonymous substitutions to the number of synonymous substitutions will

be different for the different codes. In symmetric graphs, the amino acid changing (non-synonymous) and amino acid conserving (synonymous) nucleotide sites might evolve at similar rates if they pertain to the same orbit.

In asymmetric graphs, the choice of an amino acid is not enslaved to the choice of any other amino acid. The asymmetry observed is an ancient property that left intact the universality of the SGC and the selective forces that shaped the evolving codes. The asymmetry was frozen since at least the Extended RNA codes [7]. The SGC terminates its influence after the aminoacylation of each tRNA. Asymmetry of PHGs means that amino acids have more degrees of freedom for exploring the sequence space for innovation in protein evolution. Afterwards, the asymmetry of PHGs of codons and anticodons is no longer influenced by the SGC. Asymmetry of the PHGs of codons and anticodons have facilitated the astonishing diversity of living organisms. We contend that several symmetrical codes were formed during evolution, but only the one(s) that had asymmetrical PHGs of codons and anticodons prevailed. The lack of symmetry in the PHGs of the SGC as compared to the ones of the synthetic codes shows that the usual SGC table is visually well-organized by grouping the amino acids in boxes, where the distribution of such boxes is not symmetric. In addition, the organization of amino acids in boxes is given by the wobble effect and the degeneration of the SGC. The result of the stationary distribution from the SGC and the synthetic codes shows that if the distribution of the amino acids in the SGC, i.e., the boxes of the amino acids was symmetric, the frequency of occurrence of different amino acids would not be independent of each other given neutral point mutations. In fact, the implementation of the genetic code would be biased to maintain similar frequencies of the amino acids with the same codonicity. This relation of symmetry/asymmetry describes two features of the genetic code. The symmetric properties of the SGC allow for robustness, for it to be less error-prone in its structure, whereas the asymmetric feature grants independence to the amino acids in protein evolution. The codonicity of an amino acid affects its codon usage bias which has been shown to be in co-evolution with the tRNA gene composition and it is in agreement with the selection-mutation-drift theory of codon usage in the optimization of translation [35].

In comparison with the SGC, the symmetry of the PHGs is broken by the absence of 5' A anticodons and of the anticodons that correspond to the stop codons. Note that sy2-codes have the same degeneracy of the SGC and yet some of them displayed symmetrical graphs. Therefore, degeneracy and asymmetry work together to achieve the optimality of the SGC [36]. The PHGs of codons and anticodons of mitochondrial codes turned out to be also asymmetric (not shown) [33]. Symmetry allows regularities but asymmetries allow evolvability. SGC is the result of an orchestrated coevolution of several molecules as mentioned in introduction. Life possess the salient feature of being formed by a plethora of codes that operate at different scales.

A review of group theory and abstract algebra applied to molecular systems biology can be found in [37]. These theoretical approaches can enlighten complex problems in biology. Mathematics in biology has often been regarded as an intruder perhaps for its descriptive character of the latter. Yet, it is necessary to take cognizance of the fact that most biological signals and biological systems are complex. Mathematics has become pervasive in all areas in biology. Collaboration through interdisciplinary analyses can provide new insights to complex problems such as the origin and evolution of the SGC and proteins.

Supplementary Materials: The Supplementary Materials are available online at <http://www.mdpi.com/2073-8994/12/6/997/s1>.

Author Contributions: M.V.J. and G.S.Z. conceived the whole work, contributed with ideas; M.V.J. and G.S.Z. performed the analyses; M.V.J. coordinated the research. M.V.J. and G.S.Z. wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: M.V.J. was funded by Dirección General de Asuntos del Personal Académico (DGAPA), Universidad Nacional Autónoma de México, UNAM (PAPIIT-IN201019); G.S.Z. is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and received a doctoral fellowship from CONACYT (number: 737920).

Acknowledgments: We thank Francisco Prosdocimi for his critical comments and Juan R. Bobadilla for material support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nirenberg, M.W.; Matthaei, J.H. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. USA* **1961**, *47*, 1588–1602. [[CrossRef](#)]
2. Nirenberg, M.W.; Jones, O.W.; Leder, P.; Clark, B.F.C.; Sly, W.S.; Pestka, S. On the Coding of Genetic Information. *Cold Spring Harb. Symp. Quant. Biol.* **1963**, *28*, 549–557. [[CrossRef](#)]
3. Gamow, G. Possible Relation between Deoxyribonucleic Acid and Protein Structures. *Nature* **1954**, *173*, 318. [[CrossRef](#)]
4. Crick, F.H.C.; Griffith, J.S.; Orgel, L.E. Codes Without Commas. *Proc. Natl. Acad. Sci. USA* **1957**, *43*, 416–421. [[CrossRef](#)]
5. Hornos, J.E.M.; Hornos, Y.M.M. Algebraic model for the evolution of the genetic code. *Phys. Rev. Lett.* **1993**, *71*, 4401–4404. [[CrossRef](#)]
6. Sánchez, R.; Morgado, E.R.; Grau, R. A genetic code Boolean structure. I. The meaning of Boolean deductions. *Bull. Math. Biol.* **2005**, *67*, 1–14.
7. José, M.V.; Morgado, E.R.; Govezensky, T. An extended RNA code and its relationship to the standard genetic code: An algebraic and geometrical approach. *Bull. Math. Biol.* **2007**, *69*, 215–243. [[CrossRef](#)]
8. Jiménez-Montaño, M.A.; De La Mora-Basáñez, C.R.; Pöschel, T. The hypercube structure of the genetic code explains conservative and non-conservative aminoacid substitutions in vivo and in vitro. *BioSystems* **1996**, *39*, 117–125. [[CrossRef](#)]
9. Eriani, G.; Delarue, M.; Poch, O.; Gangloff, J.; Moras, D. Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. *Lett. Nat.* **1990**, *347*, 203–206. [[CrossRef](#)]
10. Tamura, K. Origins and Early Evolution of the tRNA Molecule. *Life* **2015**, *5*, 1687–1699. [[CrossRef](#)]
11. Kim, Y.; Opron, K.; Burton, Z.F. A tRNA- and anticodon-centric view of the evolution of aminoacyl-trna synthetases, trnaomes, and the genetic code. *Life* **2019**, *9*, 37. [[CrossRef](#)]
12. Lei, L.; Burton, Z.F. Evolution of life on earth: TRNA, aminoacyl-tRNA synthetases and the genetic code. *Life* **2020**, *10*, 21. [[CrossRef](#)]
13. Koonin, E.V.; Novozhilov, A.S. Origin and Evolution of the Universal Genetic Code. *Annu. Rev. Genet.* **2017**, *51*, 45–62. [[CrossRef](#)]
14. José, M.V.; Morgado, E.R.; Govezensky, T. Genetic hotels for the standard genetic code: Evolutionary analysis based upon novel three-dimensional algebraic models. *Bull. Math. Biol.* **2011**, *73*, 1443–1476. [[CrossRef](#)]
15. José, M.V.; Zamudio, G.S.; Morgado, E.R. A unified model of the standard genetic code. *R. Soc. Open Sci.* **2017**, *4*, 160908. [[CrossRef](#)]
16. Guimarães, R.C.; Moreira, C.H.C.; de Fariás, S.T. A self-referential model for the formation of the genetic code. *Theory Biosci.* **2008**, *127*, 249–270. [[CrossRef](#)]
17. Wong, J.T.-F.; Ng, S.-K.; Mat, W.-K.; Hu, T.; Xue, H. Coevolution Theory of the Genetic Code at Age Forty: Pathway to Translation and Synthetic Life. *Life* **2016**, *6*, 12. [[CrossRef](#)]
18. Wong, J.T.-F. Coevolution theory of the genetic code at age thirty. *BioEssays* **2005**, *27*, 416–425. [[CrossRef](#)]
19. Di Giulio, M. The coevolution theory of the origin of the genetic code. *Phys. Life Rev.* **2004**, *1*, 128–137. [[CrossRef](#)]
20. Wong, J.T.-F. A Co-Evolution Theory of the Genetic Code. *Proc. Natl. Acad. Sci. USA* **1975**, *72*, 1909–1912. [[CrossRef](#)]
21. Trifonov, E.N. Consensus temporal order of amino acids and evolution of the triplet code. *Gene* **2000**, *261*, 139–151. [[CrossRef](#)]
22. Ilardo, M.; Bose, R.; Meringer, M.; Rasulev, B.; Grefenstette, N.; Stephenson, J.; Freeland, S.J.; Gillams, R.J.; Butch, C.J.; Cleaves, H.J. Adaptive Properties of the Genetically Encoded Amino Acid Alphabet Are Inherited from Its Subsets. *Sci. Rep.* **2019**, *9*, 1–9. [[CrossRef](#)]
23. Freeland, S.J.; Wu, T.; Keulmann, N. The case for an error minimizing standard genetic code. *Orig. Life Evol. Biosph.* **2003**, *33*, 457–477. [[CrossRef](#)]

24. José, M.V.; Govezensky, T.; García, J.A.; Bobadilla, J.R. On the evolution of the standard genetic code: Vestiges of critical scale invariance from the RNA world in current prokaryote genomes. *PLoS ONE* **2009**, *4*, e4340. [[CrossRef](#)]
25. Carter, C.W.; Wills, P.R. Experimental solutions to problems defining the origin of codon-directed protein synthesis. *BioSystems* **2019**, *183*, 103979. [[CrossRef](#)]
26. Crick, F.H.C. The origin of the genetic code. *J. Mol. Biol.* **1968**, *38*, 367–379. [[CrossRef](#)]
27. José, M.V.; Morgado, E.R.; Guimarães, R.C.; Zamudio, G.S.; de Fariás, S.T.; Bobadilla, J.R.; Sosa, D. Three-Dimensional Algebraic Models of the tRNA Code and 12 Graphs for Representing the Amino Acids. *Life* **2014**, *4*, 341–373. [[CrossRef](#)]
28. Zamudio, G.S.; José, M.V. Identity Elements of tRNA as Derived from Information Analysis. *Orig. Life Evol. Biosph.* **2018**, *48*, 73–81. [[CrossRef](#)]
29. Zamudio, G.S.; Palacios-Pérez, M.; José, M.V. Information theory unveils the evolution of tRNA identity elements in the three domains of life. *Theory Biosci.* **2020**, *139*, 77–85. [[CrossRef](#)]
30. José, M.V.; Morgado, E.R.; Sanchez, R.; Govezensky, T. The 24 possible algebraic representations of the standard genetic code in six or in three dimensions. *Adv. Stud. Biol.* **2012**, *4*, 119–152.
31. Zamudio, G.S.; Prosdocimi, F.; de Fariás, S.T.; José, M.V. A neutral evolution test derived from a theoretical amino acid substitution model. *J. Theor. Biol.* **2019**, *467*, 31–38. [[CrossRef](#)]
32. Zamudio, G.S.; José, M.V. On the Uniqueness of the Standard Genetic Code. *Life* **2017**, *7*, 7. [[CrossRef](#)]
33. José, M.V.; Zamudio, G.S. Symmetrical properties of graph representations of genetic codes: From genotype to phenotype. *Symmetry* **2018**, *10*, 388. [[CrossRef](#)]
34. José, M.V.; Zamudio, G.S.; Palacios-Pérez, M.; Bobadilla, J.R.; de Fariás, S.T. Symmetrical and Thermodynamic Properties of Phenotypic Graphs of Amino Acids Encoded by the Primeval RNY Code. *Orig. Life Evol. Biosph.* **2015**, *45*, 77–83. [[CrossRef](#)]
35. Rocha, E.P.C. Codon Usage Bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.* **2004**, *14*, 2279–2286. [[CrossRef](#)]
36. Zamudio, G.S.; José, M.V. Phenotypic Graphs and Evolution Unfold the Standard Genetic Code as the Optimal. *Orig. Life Evol. Biosph.* **2018**, *48*, 83–91. [[CrossRef](#)]
37. Rietman, E.A.; Karp, R.L.; Tuszynski, J.A. Review and application of group theory to molecular systems biology. *Theor. Biol. Med. Model.* **2011**, *8*, 21. [[CrossRef](#)]






© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Copyright of Symmetry (20738994) is the property of MDPI Publishing and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.

Apéndice 9

Article

The Ancient History of Peptidyl Transferase Center Formation as Told by Conservation and Information Analyses

Francisco Prosdocimi ^{1,2,*} , Gabriel S. Zamudio ² , Miryam Palacios-Pérez ²,
Sávio Torres de Farias ³ and Marco V. José ^{2,*} 

¹ Laboratório de Biologia Teórica e de Sistemas, Instituto de Bioquímica Médica Leopoldo de Meis, Universidade Federal do Rio de Janeiro, Rio de Janeiro 21.941-902, Brazil

² Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, Ciudad Universitaria, CDMX 04510, Mexico; gazaso92@gmail.com (G.S.Z.); mir.pape@iibiomedicas.unam.mx (M.P.-P.)

³ Laboratório de Genética Evolutiva Paulo Leminsk, Departamento de Biologia Molecular, Universidade Federal da Paraíba, João Pessoa, Paraíba 58051-900, Brazil; stfarias@yahoo.com.br

* Correspondence: prosdocimi@bioqmed.ufrj.br (F.P.); marcojose@biomedicas.unam.mx (M.V.J.)

Received: 7 July 2020; Accepted: 31 July 2020; Published: 5 August 2020



Abstract: The peptidyl transferase center (PTC) is the catalytic center of the ribosome and forms part of the 23S ribosomal RNA. The PTC has been recognized as the earliest ribosomal part and its origins embodied the First Universal Common Ancestor (FUCA). The PTC is frequently assumed to be highly conserved along all living beings. In this work, we posed the following questions: (i) How many 100% conserved bases can be found in the PTC? (ii) Is it possible to identify clusters of informationally linked nucleotides along its sequence? (iii) Can we propose how the PTC was formed? (iv) How does sequence conservation reflect on the secondary and tertiary structures of the PTC? Aiming to answer these questions, all available complete sequences of 23S ribosomal RNA from Bacteria and Archaea deposited on GenBank database were downloaded. Using a sequence bait of 179 bp from the PTC of *Thermus thermophilus*, we performed an optimum pairwise alignment to retrieve the PTC region from 1424 filtered 23S rRNA sequences. These PTC sequences were multiply aligned, and the conserved regions were assigned and observed along the primary, secondary, and tertiary structures. The PTC structure was observed to be more highly conserved close to the adenine located at the catalytical site. Clusters of interrelated, co-evolving nucleotides reinforce previous assumptions that the PTC was formed by the concatenation of proto-tRNAs and important residues responsible for its assembly were identified. The observed sequence variation does not seem to significantly affect the 3D structure of the PTC ribozyme.

Keywords: peptidyl transferase center; origin of life; 23S rRNA; proto-tRNA; emergence of biological systems

1. Introduction

The peptidyl transferase center (PTC) is the catalytic center of the ribosome. Being a specific region of the larger ribosomal subunit, it is responsible for binding activated amino acids together and performing peptide elongation during protein synthesis. Since the early 1980s, Carl Woese and Harry Noller noticed that the essential mechanism underlying translation might be RNA based [1]. Nevertheless, it was only in 1992 that Noller and his collaborators found experimental evidence to support the idea that the PTC was indeed a ribozyme. They confirmed that the activity of peptidyl transferase is held by the ribosomal RNA after treating ribosomes with proteases without prejudice

to the peptidic bond formation [2]. Four years later, Peter Lohse and Jack Szostak carried out the in vitro selection of ribozymes with the capability “to synthesize ester and amide linkages, as does the ribosomal peptidyl transferase” [3]. Further studies confirmed that the ribosomal peptidyl transferase reaction was performed by a region smaller than 200 base pairs located in the 23S ribosomal RNA of prokaryotes. Eukaryotes contain a similar PTC located in their 28S ribosomal RNA.

The PTC region has been considered crucial in the understanding about the origins of life. It has been described as the most significant trigger that engendered a mutualistic behavior between nucleic acids and peptides, allowing the emergence of biological systems [4–6]. Additionally, the proposal of a First Universal Common Ancestor (FUCA) departed from the contingent appearance of an ancient ribozyme capable of binding amino acids together [7]. The emergence of this proto-PTC is a prerequisite to couple a chemical symbiosis between RNAs and peptides that further evolved both to (i) become the large subunit of the ribosome by the principle of accretion and (ii) to allow the emergence of the genetic code. Although there remains controversy in the literature about whether the PTC is ancient or not [8–11], its importance cannot be challenged as it composes the central core of the decoding language of biology. The PTC is a versatile catalyst [12] that works as a turnstile for binding 20 different and very specific L-amino acids together to compose every cellular protein [13,14].

The origin and initial evolution of the PTC is a fertile field of debate and discussion in the scientific community. Some works indicated that the PTC was formed by a duplication of ancient forms of RNA once its structure was symmetric [15,16]. Other studies proposed the formation of the PTC by the junction of smaller RNAs, such as primitive tRNAs. Tamura [17] analyzed the secondary structure of the PTC and observed topological similarities with the secondary structure of transfer RNAs. In this sense, Caetano-Anollés and Sun [18] used structural analyses to provide evidence that tRNAs were older than ribosomes and were coopted to operate in the translation machinery. Farias and collaborators [19] analyzed sequence similarities between reconstructed ancestral sequences of tRNAs and the PTC. They verified an identity of 50.5% between a modern PTC and concatemers of ancestral tRNAs. Additionally, Root-Bernstein and Root-Bernstein [20] studied the similarity between tRNAs and rRNAs from *Escherichia*, observing several tRNA sequences found along its 23S rRNA sequence. They also suggested that the ribosomal RNA might have functioned as a primitive genome. Farias et al. [21] reconstructed a 3D structure of the PTC based on an ancestral sequence of tRNAs and observed a structural similarity of 92% when compared to the PTC of the bacteria *Thermus thermophilus*. Additionally, Demongeot and Seligmann [22] performed comparative studies between the secondary structure of both tRNAs and rRNAs and suggested that rRNAs were probably originated from tRNA molecules. Together, all these data make evident a scenario for the origin of life in which an evolutionary and chronological connection can be observed between these two essential components of the translation system: tRNAs and rRNAs.

Due to its remarkable relevance to biology and to the origins of life field, new studies that approach issues relating tRNAs and rRNAs are indispensable to better clarify how the initial organization of biological systems took place. Even when most works about ribosomal structure indicate that the PTC is highly conserved among all forms of life, we were unable to find conservation analyses of this particularly interesting region of the ribosome among the ancient domains of life. In addition, it seems important to analyze both the sequence and structure of the PTC in detail to gain insights about the emergence of biological systems. In this work, the following questions were posited: (i) Which exact nucleotides from the PTC are conserved among prokaryotes? (ii) How was the PTC probably formed? (iii) How can molecular modeling answer questions about the 3D structure conservation of the catalytic site of the ribosome? (iv) Are there co-evolving clusters of nucleotides that were invariant throughout the PTC's evolution? Herein, we used comparative genomics and information theory to unravel patterns of information variation and nucleotide conservation among PTCs using all complete sequences of 23S rRNAs available in the GenBank database [23].

2. Material and Methods

2.1. Download of Complete 23S Ribosomal RNAs from Public Databases

All available sequences (complete) of 23S rRNA were retrieved from GenBank using the following search: “23s ribosomal RNA [All Fields] AND complete [All Fields] AND biomol_rrna [PROP]” with the nucleotide search function of the National Center for Biotechnology Information (NCBI) website. This search resulted in 1434 sequences downloaded from GenBank [23].

2.2. Retrieving PTCs from 23S Ribosomal RNA Sequences

A PTC sequence containing 179 bp from the bacteria *T. thermophilus* was obtained [19] and used as bait to retrieve the PTC region from the other 23S rRNA sequences obtained. The selection of PTC regions was performed using the optimal pairwise alignment tool Needleman–Wunsch [24]. Each of the 1434 23S rRNA sequences downloaded were aligned to the PTC of *T. thermophilus* using the needle script provided in the EMBOSS package [25]. An in-house needleParser.pl Perl script was developed to retrieve the start and end coordinates of the alignment and another script named get_RegionByCoordinate.pl was used to retrieve the exact PTC from the 23S rRNA sequences obtained.

2.3. Multiple Alignment, Filtering, and Production of PTC Datasets

The 1434 PTC regions were multiple aligned using ClustalW [26] software. After visual inspection of the alignment, we noticed 10 sequences particularly divergent from the others, presenting exceeded nucleotides and possibly representing annotation errors. These sequences were filtered to provide a PTC dataset in FASTA format containing 1424 high-quality PTC regions.

The whole PTC dataset was also separated into three different subsampled datasets according to taxonomic information. We analyzed the whole dataset and sequences from two ancient domains of life: Bacteria and Archaea. Five PTC sequences from eukaryotic organisms were discarded from further analyses as it has been found that eukaryotes compose a derived clade originated from the Lokiarchaeota group of archaea from the subphylum Asgard [27,28]. Each domain dataset was analyzed separately. The conservation analysis of nucleotides was performed using sequence alignments and the WebLogo tool [29] was used to generate pictures identifying the most conserved nucleotide residues. Manual curation was also performed in the alignments in order to allow the identification of 100% conserved nucleotides.

2.4. Information Theory Analysis of PTCs

For the information theory analysis, we considered the pseudometric variation of information. Pseudometrics differ from metrics because two different elements can be at distance zero. The variation of information measures the distance between two messages, X and Y . This metric is given by the formula $V(X,Y) = H(X) + H(Y) - 2I(X,Y)$, where $H(X)$ stands for the entropy of the random variable X and $I(X,Y)$ is the mutual information shared between the two random variables X and Y [30,31]; both measures are considered in bit units. For the analysis of PTC sequences, we deduced from the alignment the discrete distribution of nucleotides at each position. This procedure allowed us to determine the information distance of any two sites on any set of sequences of the same length [30,32]. If two positions are at an information distance of 0, the occurrences of nucleotides at these positions are strictly predictable, i.e., it is possible to determine one nucleotide from the other. Note that this fact holds in conserved positions but also in the case when the sites present some sort of linked variation. The variation of information allowed us to cluster sites according to the information distance between them. Particularly, an intra-cluster information distance of 0 provided well-defined, non-fuzzy clusters, on which nucleotides within single clusters perfectly co-varied.

2.5. Mapping Conservation into 2D and 3D PTC Structures

The 2D structure of the PTC was obtained from Ribovision [33]. We downloaded the SVG picture from the large subunit of *T. thermophilus* and cut the region previously identified as the PTC. The picture was edited by hand using image editors. Plus, a predicted secondary structure for the PTC of *T. thermophilus* (comprising 179 bp) was generated using the software RNAstructure [34] and visualized in the foRNA applet [35]. Both structures were colored according to the variation of information of the PTC alignment of all the 1425 sequences. Both the sequence and the 3D structure of *T. thermophilus* 23S rRNA (PDB ID 4v4i) were downloaded and manually edited to obtain the PTC region only.

2.6. 3D Modeling of the Different PTC Sequences and Structural Comparisons

Using the UGENE software [36], consensus sequences were obtained from each alignment file. The ModeRNA webserver [37] was used to perform template-based 3D modeling using the *T. thermophilus* PTC as a model to predict the 3D structure of the consensus sequences from the three datasets under analysis. The modeled structures were structurally aligned using the RNA-align software from the Zhang Lab suite [38]. Finally, we used Chimera [39] to visualize the structural comparisons.

3. Results

3.1. PTC Datasets: Production and Taxonomic Analysis

On 7 October 2019, the GenBank database contained 1434 complete sequences of 23S ribosomal RNAs. All these sequences were downloaded in FASTA format and aligned (using an optimal pairwise alignment tool) to the PTC of the bacteria *T. thermophilus* to map a PTC sequence region containing 179 bp. Using in-house Perl scripts, we parsed the PTC alignment information and generated a file containing 1434 PTC sequences. After an initial round of multiple alignment of the PTC dataset, we visually identified ten sequences that seemed too divergent in the alignments. These anomalous PTC sequences were removed from further analyses as they possibly represented sequences with inaccurate genome annotation [40]. Therefore, a dataset containing 1424 PTC sequences in FASTA format was produced. This dataset presented five sequences from eukaryotes, 118 sequences from archaea and 1301 sequences from bacteria. Inside the bacterial clade, we observed that the major groups sampled were Proteobacteria (564 sequences), Firmicutes (237 sequences), and Actinobacteria (153 sequences); another 347 sequences were divided into 25 other clades. From the Archaea domain, 81 sequences came from Euarchoeota, 33 from Crenarchaeota, three from Thaumarchaeota, and one from Nanoarchaeota. The five sequences from eukaryotes came from the fungi species *Encephalitozoon intestinalis*. Table 1 summarizes the information about the sequences obtained.

Table 1. Peptidyl transferase center (PTC) sequences per clade, number of sequences, and multiple alignment features.

Dataset Name	Clades	Number of PTC Sequences	Alignment Size (Gaps)	Positions 100% Conserved
PTC-all	Bacteria	1424	201 (10)	42
	Archaea			
	Eukarya			
PTC-Bac	Bacteria *	1301	195 (8)	62
PTC-Arc	Archaea *	118	186 (7)	83
PTC-Pro	Proteobacteria *	564	186 (7)	110
PTC-Fir	Firmicutes **	237	184 (4)	122
PTC-Act	Actinobacteria	153	179 (0)	132

* These datasets are subsets of the PTC-all dataset. ** These datasets are subsets of the PTC-Bac dataset.

3.2. Multiple Alignment and Sequence Conservation of PTC Datasets

A multiple alignment of each PTC dataset was performed and followed by the analysis of sequence conservation in the main datasets. Table 1 shows that all dataset alignments were 179 bp in length plus the number of gaps added by the alignment tool to optimize the sequence alignment. As expected, due to the presence of highly divergent sequences, the dataset PTC-all presented the highest number of gaps (10) and the fewest number of 100% conserved nucleotides (42). The highest number of positions 100% conserved was found in the Actinobacteria dataset, with 73.7% of conserved nucleotides and no gaps found. This was followed by Firmicutes, with 66.3% of residues completely conserved, and Proteobacteria, with 59.1%, evidencing a higher diversity of PTC sequences in the latter clade. This possibly happened due to the existence of large sub-clades (such as Alpha-, Beta, Gamma-, and Deltaproteobacteria), presenting well-differentiated PTC sequences among them.

The multiple alignments of the three main PTC datasets, i.e., PTC-all, PTC-Bac, and PTC-Arc, are displayed as sequence logos (Figure 1), on which the conservation profile can be easily visualized. In sequence logos, the 100% conserved positions were shaded in green, and the main adenine located at the catalytic site was starred on top. The observed single nucleotide gaps reveal variability in size among sequences. In some cases, the difference was detected in a single nucleotide present in one individual sequence, such as (i) the gap observed in position 21 of the PTC-all dataset; (ii) position 169 in PTC-all (and its equivalent position 165 in PTC-Bac) appeared merely due to the presence of a G in the bacterium *Spirochaeta africana*; and (iii) position 16 in PTC-Bac that was represented due to a C observed solely in the bacterium *Streptococcus pyogenes*. Other exceptional cases of gaps include positions 106 and 120 in PTC-all (equivalent to positions 101 and 115 in PTC-Bac) that correspond to single nucleotides observed in two Gammaproteobacteria species. In other cases, the differences rely on a few sequences that belong to one specific clade; for example, the gap in position 53 in PTC-all and position 49 in PTC-Bac correspond to a T present in Firmicutes species.

The lowest nucleotide conservation observed in the PTC sequences spans from nucleotide 37 to 44 and 65 to 70 in PTC-all and happened due to the fact that several bacteria from the phyla Chloroflexi, Aquificae, Fusobacteria, Bacteroidetes, Thermotogae, and Elusimicrobia (as well as some archaea) present divergent nucleotides in that region. In PTC-Bac, the sections ranging from 32 to 41 and 62 to 66 are observed to be more variable because the sequence data for the aforementioned phyla do not have analogous counterparts in archaea.

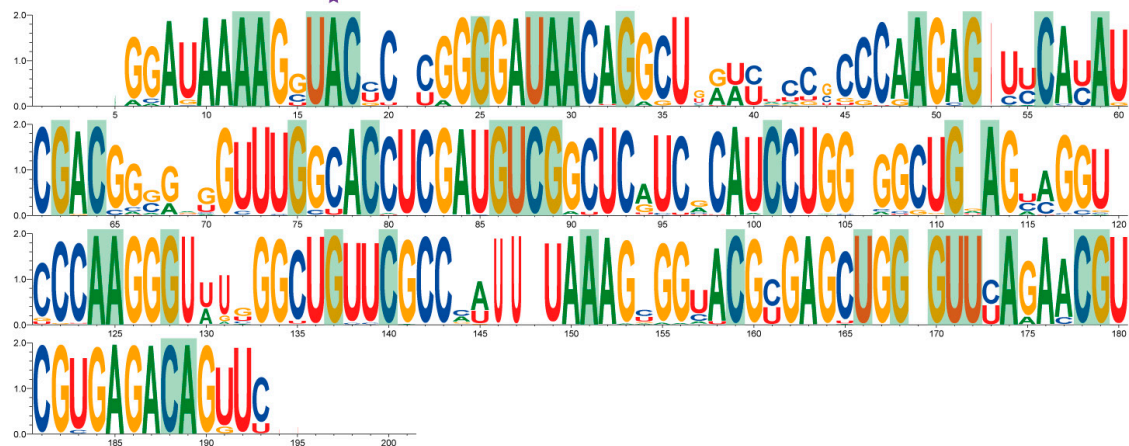
Considering the gaps spanning more than one nucleotide, it is possible to see a void at the beginning of the PTC-all alignment. It corresponds to a certain variation observed in the 5' PTC region of some euryarchaeal sequences that seem to present a duplication in their five initial nucleotides. Such a region has a more notorious representation in the PTC-Arc dataset. The gap observed at the 3' end of the PTC-all alignments and nearly at the end of the PTC-Bac alignments appeared due to one single sequence coming from the cyanobacterium *Thermosynechococcus elongatus* that differs from all other organisms. Therefore, the multiple alignment tool positioned the sequences in the most convenient way, either keeping the gap at the very end of the PTC-all dataset alignment (from nucleotides 194 to 201) or introducing a gap just before the end, as observed in PTC-Bac (between sites 181 and 185).

Interestingly, the only gaps observed in the PTC-Arc dataset appear at both ends (5' and 3'). The initial one corresponds to the previously mentioned apparent duplication of the initial nucleotides in some euryarchaeal sequences. There is also a short portion at the last two nucleotides of the PTC-Arc dataset that shape an apparent gap at the end of alignment. This is observed because the PTC from *Nanoarchaeum equitans* and from some crenarchaeal sequences have a couple of G nucleotides inserted there. Additionally, the PTC-Arc sequence logo (Figure 1c) shows the highest amount of heterogeneity among all three datasets, observed in sites with different sizes of nucleotides along the vertical axis. Notably, even if the PTC-Arc dataset presented the highest number of conserved sites along the two domains analyzed (83 sites shaded in green for archaea compared to 62 for bacteria), its non-conserved

its non-conserved sites also displayed higher variation. This fact denotes both the conservation of a tight PTC structure and a significant variability of this diverse clade that originated eukaryotic organisms [27,28].

Finally, site A2451 from the 23S rRNA is the catalytic site of the PTC, essential for the peptide bond to occur. This nucleotide has been shown to be absolutely preserved in each and every analyzed sequence, as it can be observed at the highlighted site with a magenta star at A179 in PTC-all, A13 in PTC-Bac, and A17 in PTC-Arc (Figure 1).

a) PTC-all



b) PTC-Bac



c) PTC-Arc

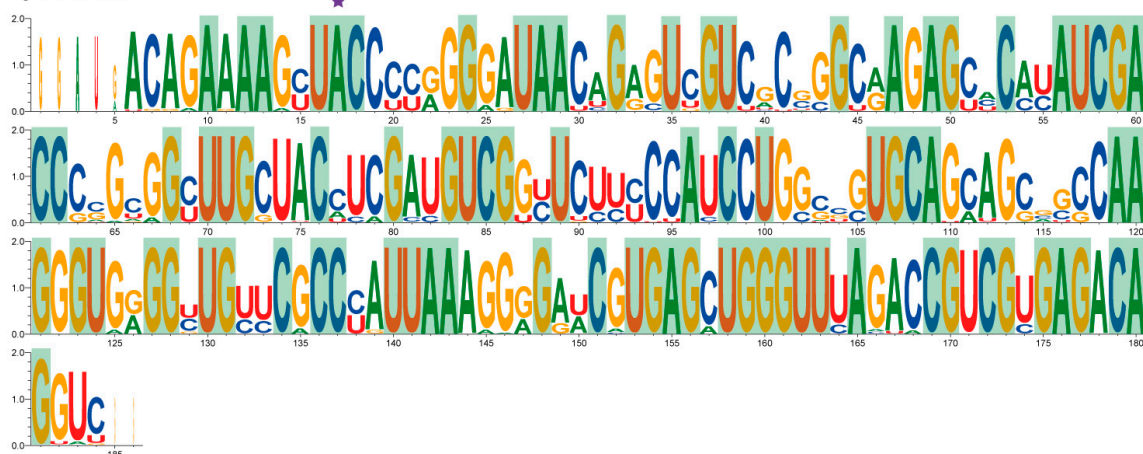


Figure 1. WebLogos showing nucleotide conservation in the main analyzed PTC datasets. (a) PTC-all; (b) PTC-Bac; (c) PTC-Arc. Universally conserved nucleotides in each dataset are shown with a green background. The adenine located at the catalytic site is highlighted with a magenta star.

Life 2020, 10, × FOR PEER REVIEW

7 of 16

Finally, site A2451 from the 23S rRNA is the catalytic site of the PTC, essential for the peptide bond (Figure 1). This nucleotide has been shown to be absolutely conserved in each and every analyzed sequence, by PTC-Bac (observed at the highlighted site with a magenta star) and A17 in PTC-Arc, A13 in PTC-Bac, and A17 in PTC-Arc (Figure 1a). The catalytic site A2451 highlighted with a magenta star.

3.3.3 Mapping 100% conserved sites into the 3D structure of the PTC

To gain insights about the nucleotide conservation in the PTC, we produced 3D models in which the 100% conserved positions (shaded green in Figure 1 and drawn in black in Figure 2) could be seen over the tridimensional structure. Thus, we obtained PTC consensus sequences for each dataset and modeled their structures using MolProbity software using the known PTC structure of *Thermoplasma* (PDB ID: 1V4H) as a template. Analyzing Figure 2, we observe a higher number of conserved positions aggregated at the top of the structure (shaded oval in Figure 2), close to the catalytic site A2451 (identified with a magenta star). Nevertheless, the entire structure is significantly conserved, and specific conserved nucleotides located at different sites along the whole structure are probably anchors for holding the 3D shape of the PTC.

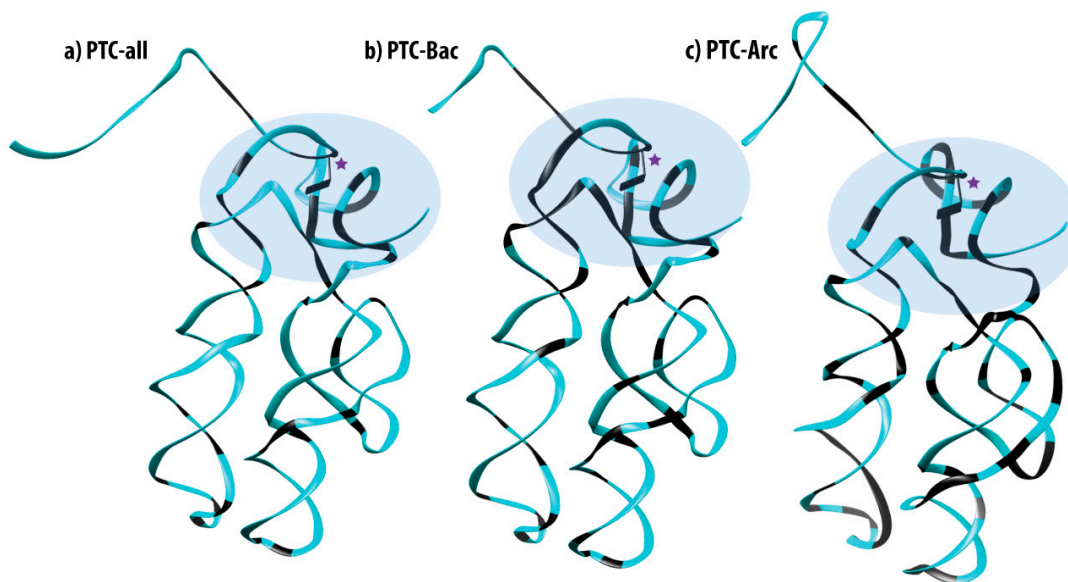


Figure 2. Embedding the 100% conserved nucleotides for each dataset (a) PTC-all (b) PTC-Bac and (c) PTC-Arc. The structures were reproduced by template-based modeling over the known *T. thermophilus* structure. The catalytic site A2451 is depicted as a magenta star. The 100% conserved nucleotides are colored in black and a blue oval highlights the highly conserved sites on the top of the structures.

3.4. Identity Elements, Entropy Variation, Information Variation

3.4.1 Identity Elements, Entropy Variation, Information Variation

For the complete set of sequence alignments, an entropy value was determined for each position. Thus, nucleotides were grouped into information clusters that were identified by color codes along the sequences (Figure S1 in Supplementary). The first graph (Figure S1a) corresponds to the analysis of both bacterial and archaeal sequence alignments, while Figure S1b,c present the data for bacterial and archaeal sequences, respectively. Given that the number of sequences used for the informational and archaeal sequences, respectively. Given that the number of sequences used for the informational analysis decreased between the PTC-all dataset and the archaeal one, the number of information clusters and positions in the clusters increased due to the reduction in variation. Ungapped positions, such as 45, 73, and 116 observed in the PTC-all dataset, have shown entropy close to the maximum value of two, meaning that all four nucleotides occurred in almost equiprobable amounts. Colored clusters with entropy equal to zero represent invariant nucleotide positions (Figure S1); while clusters with entropy greater than zero reflect positions on which nucleotide variations are highly predictable.

This property can be clearly noticeable in the entropy profile for archaeal sequences (Figure S1c) in which the positions in the red-colored cluster have entropy greater than 0. In each plot of Figure S1,

and three positions each. The information cluster harboring three positions was the only one in PTC-all whose bases were invariant, i.e., showed entropy equal to zero. The PTC-Bac dataset presented 71 positions divided into 11 clusters: Five clusters contained two positions, whereas the others presented 20, 11, 10, 8, 7, and 5 nucleotide positions. The cluster with the highest number of positions (20) was the only one that presented an intra-cluster entropy of zero. Regarding the PTC-Arc dataset, 101 positions were found split into 16 information clusters: 12 clusters contained two positions and the others possessed 4, 7, 20, 7, and 3 positions. In the archaea data, the cluster containing 20 nucleotides was unique and contained an intra-cluster entropy equal to zero, representing invariant positions. The red color is associated with the modal cluster, i.e. in Supplementary Table 1. The highest number of interdependent, co-evolving nucleotide positions.

3.5. Mapping the data shows 25 information clusters in the 2D structure of the PTC

Six of these clusters contained two positions and the remaining three clusters contained six, four, and three positions each. A subtle variation was produced for each dataset alignment to consider only their sequence positions without gaps, so that the length matched the canonical 179 nucleotides from the PTC. The information cluster harboring three positions was the only one in PTC-all whose bases were invariant, i.e., showed entropy equal to zero. The PTC-Bac dataset presented 71 positions divided into 11 clusters: Five clusters contained two positions, whereas the others presented 20, 11, 10, 8, 7, and 5 nucleotide positions. The cluster with the highest number of positions (20) was the only one that presented an intra-cluster entropy of zero. Regarding the PTC-Arc dataset, 101 positions were found split into 16 information clusters: 12 clusters contained two positions and the others possessed 4, 7, 20, 7, and 3 positions. In the archaea data, the cluster containing 20 nucleotides was unique and contained an intra-cluster entropy equal to zero, representing invariant positions. All the nucleotide positions of the clusters are shown in Supplementary Table S1.

Analyzing Figure 3, one observes that the de novo structure differs from the modern PTC in two significant regions: (i) A stem formed on the first segment by the pairings of the bases from 9:25 (Figure 3, red star) up to 14:19 (Figure 3, blue star) and forming a loop with the segment 15-18 (Figure 3b) and (ii) the extra pairings of bases 75:159 and 76:158 (Figure 3, green star), on which the segment that joins the two coils in the secondary structure of the PTC can be seen. These three stars represent the main topological differences in the 2D structures (shown in Figure 3a,b), as all the other nucleotides are arranged similarly. The stars indicate which Watson-Crick base pairings should be broken in the lower entropy arrangement of Figure 3b to produce a modern-like PTC structure, as observed in Figure 3a. It is possible that these two RNA structures were viable and interchangeable at the origins of the PTC in prebiotic Earth, as it is known that RNA molecules can adopt different conformations [41, 43]. These alternative foldings (among other possible ones) possibly changed according to the presence of different ligands and environmental conditions [44].

Thus, information clusters were computed one more time and mapped into both (i) the known secondary structure of the PTC from *T. thermophilus* (Figure 3a) and (ii) the secondary structure predicted de novo by the software forNA, using the PTC-all consensus sequence as entry (Figure 3b). These alternative foldings (among other possible ones) possibly changed according to the presence of different ligands and environmental conditions [44].

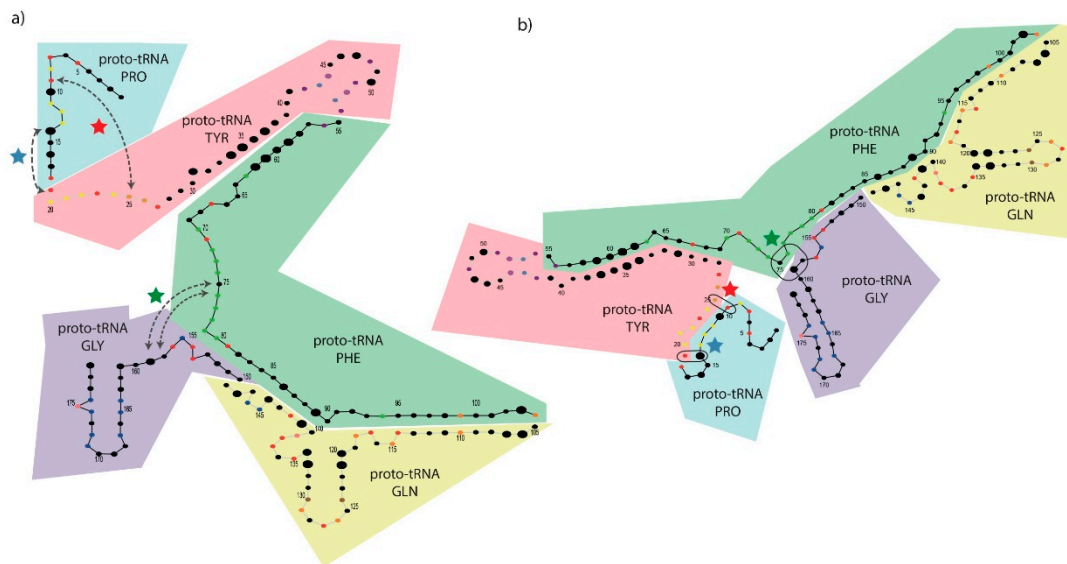


Figure 3. Information clusters and proto-tRNA composition from the PTC-Bac dataset as observed over (a) the secondary structure of the modern PTC from *T. thermophilus* and (b) the de novo (predicted) RNA structure. The radius of each circle corresponds to its entropy value, i.e., bigger circles represent more variable positions. Regions corresponding to each proto-tRNA are shown in colored boxes according to the corresponding ancestors. Colored stars in blue, red, and green represent Watson-Crick base pairing in putative ancestral folding (b) that were separated to generate the modern folding (a). Arrows in (a) represent the positions linked by Watson-Crick base pairing in (b) that were separated to produce the catalytic structure of the PTC (a).

Analyzing Figure 3, one observes that the de novo structure differs from the modern PTC in two significant regions: (i) A stem formed on the first segment by the pairings of the bases from 9:25 (Figure 3, red star) up to 14:19 (Figure 3, blue star) and forming a loop with the segment 15–18 (Figure 3b) and (ii) the extra pairings of bases 75:159 and 76:158 (Figure 3, green star), on which the segment that joins the two coils in the secondary structure of the PTC can be seen. These three stars represent the main topological differences in the 2D structures (shown in Figure 3a,b), as all the other nucleotides are arranged similarly. The stars indicate which Watson–Crick base pairings should be broken in the lower entropy arrangement of Figure 3b to produce a modern-like PTC structure, as observed in Figure 3a. It is possible that these two RNA structures were viable and interchangeable at the origins of the PTC in prebiotic Earth, as it is known that RNA molecules can adopt different conformations [41–43]. These alternative foldings (among other possible ones) possibly changed according to the presence of different ligands and environmental conditions [44].

3.6. Mapping Proto-tRNAs into the 2D Structure of the PTC

When trying to explain how the PTC might have been formed in the past, we benefited from the work of Farias [45], who obtained putative ancestral tRNAs (Supplementary Information) using a dataset of 9758 sequences downloaded from a tRNA database [46]. In that work, Farias (2013) [45] separated 22 types of tRNAs, including 20 canonical tRNAs, one initiator tRNA, and one tRNA for selenocysteine, ran ModelTest to find the best nucleotide substitution model, and produced ancestral sequences using an approach based on maximum likelihood. In a following publication, Farias and collaborators (2014) [19] used a combinatorics approach to randomly concatenate those ancestral proto-tRNAs and search for possible matches in a nucleotide alignment against protein databases. Notably, these researchers found a specific combination of five proto-tRNAs concatenated directly (+/+ strands) that was shown to present 50% nucleotide identity to the PTC of the bacterium *T. thermophilus*. Therefore, to check whether the early origin of the PTC could be explained by the concatenation of those proto-tRNAs, we took these ancestral proto-tRNAs that bound to the amino acids (i) proline (Pro), (ii) tyrosine (Tyr), (iii) phenylalanine (Phe), (iv) glutamine (Gln), and (v) glycine (Gly) and aligned them to the PTC of *T. thermophilus*. These five proto-tRNAs were therefore mapped (in the order described above) within the segments 1–18, 19–54, 55–104, 105–149, and 150–179 of the modern PTC from *T. thermophilus* and plotted in colored boxes, as illustrated in Figure 3. As observed in the original publication (Farias, 2013) [45], the ancestral proto-tRNAs presented variable sizes (as measured in base pairs) due to the variable nucleotide conservation observed in the modern tRNAs used to produce them. Therefore, the maximum likelihood model, applied by Farias (2013) [45], removed some nucleotide positions that were not shown to be conserved in most tRNAs used to build the ancestral sequences and produced a sort of “truncated” ancestral proto-tRNA. The higher the nucleotide conservation in the modern tRNA sequences (to build the ancestral sequences), the longer the proto-tRNAs were.

We proceeded to analyze the co-occurrence of those proto-tRNAs and information clusters to check whether it could provide us with some insights. To do that, we started analyzing the PTC-Bac dataset due to the fact that it presented an intermediate number of clusters, as PTC-all contained too few positions in clusters (25 nucleotides), and PTC-Arc presented too many (101 nucleotides). Thus, analyzing the bacterial dataset (containing 71 nucleotides in 11 clusters), we noticed that four out of nine information clusters contained bases corresponding to the positions located in nucleotides placed in regions mapped in two different proto-tRNAs. In particular, Figure 3 provides evidence that (i) the cluster colored in yellow contained bases (black circles) putatively coming from proto-tRNA^{Pro} and proto-tRNA^{Tyr}; (ii) the purple cluster contained bases (black circles) within proto-tRNA^{Tyr} and proto-tRNA^{Phe}; (iii) the orange cluster contained bases (black circles) found in proto-tRNA^{Phe} and proto-tRNA^{Gln}; and (iv) the dark blue cluster contained bases (black circles) found in both proto-tRNA^{Gln} and proto-tRNA^{Gly}. We hypothesize that these information clusters represent co-evolving nucleotides originally responsible for linking the proto-tRNAs together in a higher-level

secondary structure of extreme relevance to shaping the overall PTC structure. The cluster represented by (v), shown in red dots, contained bases mapped in all proto-tRNAs and was possibly relevant to the assembly of the whole 3D structure of the PTC. The one (vi) cluster shown with green dots contained a segment corresponding only to the third proto-tRNA^{Phe} (Figure 3). The other five clusters from the PTC-Bac dataset contained merely two bases representing Watson–Crick base pairs inside regions mapping single proto-tRNAs.

Similar data about the relationship between proto-tRNA mapping and information clusters are presented for PTC-all and PTC-Arc (Figure S2; Supplementary Table S1). Regarding the PTC-all dataset, we found one cluster containing six positions to have nucleotides coming from all the five proto-tRNAs. This finding reinforces the previous hypothesis of PTC formation by the concatenation of proto-tRNAs and suggests that this informational relationship might help to bind the proto-tRNAs together. Another identity cluster containing four positions was found in sites present in both proto-tRNA^{Gln} and proto-tRNA^{Gly}. The PTC-all cluster, with three positions, embraced the first, fourth, and fifth proto-tRNAs and two clusters with two bases were found to present nucleotides in two regions mapped to different proto-tRNAs. Regarding the PTC-Arc dataset, the two clusters containing the highest number of nucleotide positions (47 and 20), encompass regions mapped in all the five proto-tRNAs. The PTC-Arc cluster, containing seven nucleotides, mapped into the third, fourth, and fifth proto-tRNAs. A cluster with three positions was mapped into the first, fourth, and fifth proto-tRNAs. Finally, six out of 12 clusters, containing two positions, mapped to two different proto-tRNAs. It is also conspicuous that in both the PTC-all and PTC-Arc datasets, there exist identity clusters containing nucleotide positions shared by non-consecutive proto-tRNAs. Altogether, those mappings reinforce the hypothesis that these information clusters account for the 3D configuration of the modern PTC by linking together ancestral proto-tRNAs.

3.7. Mapping Identity Elements and Information Clusters into the 3D Structure of the PTC

In Figure 4, the tridimensional structure of the PTC derived from *T. thermophilus* 23S rRNA is shown and colored according to the different information clusters found in the PTC-Bac dataset, aiming to observe their spatial distribution. For all datasets, the most evident characteristic is that many identity clusters bundle in a tridimensional configuration (data for PTC-all and PTC-Arc are shown in S3). Besides the modal cluster (in red), most other nucleotides sharing similar clusters are observed in nearby positions. Brown sites, for example, remain together in the arm observed at the right side of the structure in Figure 4. Most of the green and dark blue nucleotides are close together in the 3D shape, near the yellow cluster that contains the catalytic site (A12 in this structure, marked with a magenta star). The red-colored nucleotides—corresponding to the cluster containing a higher number of interrelated, co-evolving positions—are spread throughout the PTC molecule, possibly responsible for maintaining the whole structure. They are absent, however, from the bottom of one arm (observed at the left side of Figure 4a) in which purple, light blue, and pink clusters mostly muster. Therefore, all clusters seem necessary to maintain the PTC structure and create the cage necessary for the peptide bonds to occur.

3.8. Structural Alignment of PTC Datasets to *T. Thermophilus*

The template-based structural reconstructions for the consensus sequences from the PTC-all, PTC-Bac, and PTC-Arc datasets were structurally compared to the actual PTC known for *T. thermophilus* in order to provide an appreciation of the spatial differences among them (Figure 5; in which cyan-colored regions represent a match between predicted and actual PTC structures and purple represents differences).

with a magenta star. The red-colored nucleotides are spread throughout the PTC molecule, possibly responsible for maintaining the whole structure. They are absent, however, from the bottom of one arm (observed at the left side of Figure 4a) in which purple, light blue, and pink clusters mostly cluster. Therefore, all clusters seem necessary to maintain the PTC structure and create the cage necessary for the peptide bonds to occur.

Figure 4. Ribbon (a) and surface (b) tridimensional representations of the PTC from *T. thermophilus* with the information clusters found in the PTC-Bac dataset colored according to Figure 3. The catalytic site (corresponding to A2451 or A12 in the current model) is marked with a magenta star and protrudes from the ribbon sketch (a).

3.8. Structural Alignment of PTC Datasets to *T. Thermophilus*

The template-based structural reconstructions for the consensus sequences from the PTC-all, PTC-Bac, and PTC-Arc datasets were structurally compared to the actual PTC known for *T. thermophilus* in order to provide an appreciation of the spatial differences among them (Figure 5; in which cyan-colored regions represent a match between predicted and actual PTC structures and purple represents differences).

As expected, due to the high general conservation shared among their sequences, the template-based structures were similar to the PTC of *T. thermophilus* [47], while the most relevant variations were observed towards their extremities. Indeed, the most similar structure to *T. thermophilus*' PTC was PTC-Bac, which was shown to be only slightly longer than the model at the 3' end of the structure. The most patent difference between the model and the consensus of PTC-all was the presumptive insertion of five nucleotides at the 3' end, whereas it was arranged as an internal insertion positioned at the top of the PTC-Arc structure.

In the portions corresponding to nucleotides 116 to 123 and 166 to 173 of the PTC from *T. thermophilus*, both PTC-all and PTC-Bac consensus differed from the model. However, the alignments revealed only a slight difference in the corresponding segments. An opposite situation happened in the comparison to the PTC-Arc structure, in which the sites 116 to 119 revealed a highly variable sequence region though the consensus structure and the models were shown to be nearly undistinguishable.

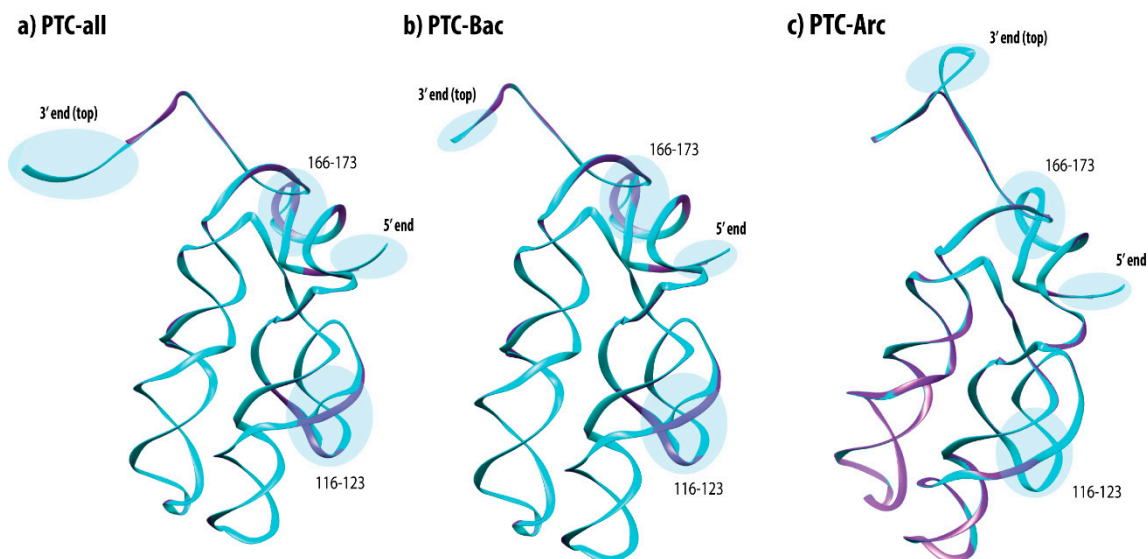


Figure 5. Tridimensional structural comparisons between the template PTC structures from *T. thermophilus* and the predicted PTCs. The template PTC structure differs from the predicted PTCs in cyan and purple. The relevant PTC regions are circled and labeled: (a) PTC-all, (b) PTC-Bac, and (c) PTC-Arc. Relevant regions are circled and labeled.

As expected, due to the high general conservation shared among their sequences, the template-based structures were similar to the PTC of *T. thermophilus* [47], while the most relevant variations were observed towards their extremities. Indeed, the most similar structure to *T. thermophilus*' PTC was PTC-Bac, which was shown to be only slightly longer than the model at the 3' end of the structure. The most patent difference between the model and the consensus of PTC-all was the presumptive insertion of five nucleotides at the 3' end, whereas it was arranged as an internal insertion in the literature about whether the ribosome was built over the PTC region or not [9, 11], many positioned at the top of the PTC-Arc structure.

In the portions corresponding to nucleotides 116 to 123 and 166 to 173 of the PTC from *T. thermophilus*, both PTC-all and PTC-Bac consensuses differed from the model. However, the alignments revealed only a slight difference in the corresponding segments. An opposite situation happened in the comparison to the PTC-Arc structure, in which the sites 116 to 119 revealed a highly variable sequence region though the consensus structure and the models were shown to be nearly undistinguishable.

4. Discussion

The sequence of the PTC is possibly the most relevant stretch of nucleic acid to be studied if one aims to understand the origin of life. Nowadays, it is a consensus that the ribosome should be understood as a prebiotic machine that predated the origin of cells [16,48]. Although there is a debate in the literature about whether the ribosome was built over the PTC region or not [9–11], many researchers claim to have evidence that the assembly of the genetic code and the ribosome started with the initial formation of region V of the ribosome, just the place in which the PTC is settled. Theoretical works on the origin of life suggest that the contingent appearance of this ribozyme capable of binding amino acids together was crucial to both the initial emergence and further development of the phenomenon of life [7,49,50].

Here, we evaluated the sequence and structural conservation of the peptidyl transferase center using all completely available sequences of 23S ribosomal RNA present in the GenBank database and annotated as such. We decided to use complete sequences to add rigor to the analyses and to avoid to the sequencing errors often present in small molecules [40]. Besides, as we were interested in understanding the relevance of the PTC to the early origin of life, we decided to exclude eukaryotic sequences from the analyses. Eukaryotes are now known to have originated from archaeal organisms coming from the phylum Lokiarchaeota, subphylum Asgard [27,28], therefore being derivate clades and having no substantial role in early origins of life.

The 1424 23S rRNA sequences obtained were aligned, filtered out to retrieve only the PTC region, and divided into three main datasets. Although there is a consensus in the literature about the fact that the PTC sequence is highly conserved [51–55], to our knowledge, PTC sequences have not been analyzed thoroughly by comparative sequence analysis, information variation, and bi/tri-dimensional structural analysis to better validate this assumption.

The multiple alignment comparison showed that the PTC from archaea presents about 40% fully conserved bases; this number lowers to 30% in bacteria, and to 20% in all analyzed organisms (Table 1). These percentages of conservation are clearly dependent on the total number of sequences analyzed, as bacteria possessed >11x more sequences available in GenBank than archaea (1302 versus 118). Curiously, the archaeal dataset presented both a higher number of conserved nucleotides (83) and a higher number of variable sites when compared to the bacterial one. This possibly means that the structure of the archaeal PTC is more optimized to be tighter in specific regions that maintain a rigid tridimensional backbone and looser in others. By contrast, the structure of the bacterial PTC is possibly wobblier.

When observed in the context of the bi- and tridimensional structures, we found that most fully conserved bases from the PTC folded close to the catalytic site, whereas sites located down to the two hairpin structures seem to allow more variation (Figure 2). This last fact should be expected, as the catalytic sites of enzymes are often more conserved than the other parts; the same is true for ribozymes. The PTC is known to be a flexible and efficient catalyst [12] as it is capable of recognizing different, specific substrates (20 different amino acids bind to aminoacyl-tRNAs) and polymerizing proteins at a similar rate [56,57]. Therefore, considering the extreme relevance of the PTC, it would be surprising if the site of catalysis showed variation.

The use of a pseudometric to show the information variation in PTC sequences allowed us to identify clusters of nucleotides that are informationally linked. We were able to find clusters containing as many as 47 nucleotides, although most of them presented fewer than 10 nucleotides. The clusters

with a higher number of bases were colored in red for all datasets and they were invariant for PTC-all and PTC-Bac, although in PTC-Arc, this modal cluster presented 47 nucleotide positions that could vary coordinately. We hypothesize that these clusters were mainly important to keep the tridimensional structure of the PTC, but we found out that they also provided interesting insights about how the PTC was formed.

Farias and collaborators [19] used tRNA sequences from hundreds of species, together with maximum likelihood analyses, to construct ancestral sequences for each of the 20 different tRNAs, producing putative ancestral proto-tRNAs. When they randomly concatenated these proto-tRNAs and BLASTed them with GenBank's nucleotide database, they verified that one concatamer of five proto-tRNAs presented a significant nucleotide identity (about 50%) to the 23S rRNA of the bacterium *T. thermophilus*, exactly in the PTC region [19]. Even if 50% identity cannot be considered a significant threshold for sequence identity, one cannot expect to apply modern standard measures of nucleotidic variation when working with an event so distant in the past. Therefore, even considering the hypothetical nature of this result, we decided to go further into that investigation. Thus, we mapped their five proto-tRNAs concatamers into the 2D structure of the *T. thermophilus* PTC. Besides mapping the proto-tRNAs, we also produced a diagram in which the nucleotides were colored according to the clusters produced with the information variation analysis. This resulted in Figure 3a, which showed the 2D structure of the PTC with dots representing each nucleotide of the PTC colored according to its corresponding cluster. We found out that many information clusters contained informationally linked nucleotides mapped to distinct proto-tRNAs along the PTC structure. This fact seems to indicate that these nucleotides may have been relevant to the stepwise binding of these ancestral tRNAs with each other in order to produce the modern shape of the PTC site. These results are in accordance with recent works suggesting that either the PTC or rRNAs should have been formed by the assembly of tRNAs [17,18,20,22]. We not only confirmed these previous assumptions but added new information on the conserved sites possibly used to link the PTC structure. Additionally, we used a de novo modeling software to predict the 2D structure of the PTC (Figure 3b) and were able to produce a structure very similar to the modern PTC. Working over this structure, we were able to identify four sites linked by Watson–Crick bonds that, when released, may have given rise to the modern PTC (these sites were identified by three colored stars in Figure 3). Our hypothesis is that the ancient PTC could be observed in at least these two structures, changing from one to the other according to the presence of ligands and specific environmental conditions—a known property of RNAs—to present multiple unstable, interchangeable structures [41,44]. Additionally, a hairpin with yellow dots observed in the bottom part of the de novo structure (Figure 3b) clearly indicated which nucleotides were used to bind proto-tRNA^{Pro} and proto-tRNA^{Tyr} into an integrated, higher-level structure that produced the PTC. Although the binding of other nucleotides between distinct proto-tRNAs cannot be clearly observed in the current structure of the PTC, we hypothesize that these co-evolving nucleotides were important to bind the proto-tRNAs together when the PTC was under formation, as its secondary structure probably grew by the addition of one tRNA at time. Thus, the informationally linked nucleotides possibly held the higher-level 2D and 3D structures together to allow the stepwise formation of the whole PTC region.

As both the position of nucleotides sharing the same cluster (Figure S1) and the bidimensional structure of the PTC (Figure 3) looked sometimes distant and peculiar, we decided to gain new insights by plotting the clusters into a tridimensional structure. We used the same color code for clusters to check whether the position of the nucleotides sharing same clusters would make more sense in 3D and we found that this was indeed the case for most clusters (Figure 4). Both the ribbon and the surface structures demonstrate that nucleotides sharing the same information cluster were usually observed close to each other in the 3D structure.

An interesting possibility derived from the current analyses would be the actual production of resurrection experiments [58] able to synthesize the putative form of an ancient PTC using the exact nucleotide sequence derived from the concatamer of proto-tRNAs described here. Which properties

may this molecule have? How would it fold? Could this sequence function as a ribozyme and catalyze a peptide bond? Similarly, it could be possible to synthesize the proto-tRNAs proposed by Farias (2013) [45] and verify whether they could bind with each other using the nucleotides described in the information variation model we used. Those experiments could bring an experimental background to the theoretical analyses performed here.

Finally, the predicted 3D structures based in the consensus sequences for each dataset were structurally compared to the known PTC structure from *T. thermophilus* [47] to allow for the identification of similarities and divergences. Despite some notable nucleotide variations from PTC-all and PTC-Bac to the *T. thermophilus* sequence, the 3D structure was shown to be significantly conserved. The higher divergences found were to be related to slight extensions observed in the 3' regions of the datasets. Regions containing nucleotide variance were shown to be conserved at the structural level (Figure 5). In conclusion, we have provided (i) a better understanding of how nucleotide variation is observed in the PTC, underscoring (ii) a testable possible model about how proto-tRNAs shaped its structure, and (iii) how the evolutionary process froze essential nucleotide positions that enabled the peptide polymerization bonding by preserving the tridimensional structure of the PTC ribozyme.

Supplementary Materials: The following are available online at <http://www.mdpi.com/2075-1729/10/8/134/s1>, Figure S1: Variation of information (bits) along the sequence of the PTC: (a) PTC-all (b) PTC-Bacteria, and (c) PTC-Archaea. Figure S2: Relationship between proto-tRNA mapping and information clusters for (a) PTC-all and (b) PTC-Arc; Figure S3: Mapping identity elements and information clusters to the 3D structure of (a) PTC-all (b) PTC-Bacteria, and (c) PTC-Archaea. Table S1: Identification of nucleotide positions clustered by the information analysis of each PTC dataset. The tRNA ancestral sequences.

Author Contributions: Conceptualization: F.P., M.V.J., S.T.d.F.; methodology: F.P., G.S.Z., M.P.-P., M.V.J.; software F.P., G.S.Z., M.P.-P.; validation: F.P., G.S.Z., M.P.-P., S.T.d.F.; formal analysis: F.P., G.S.Z., M.P.-P., S.T.d.F., M.V.J.; investigation: F.P., G.S.Z., M.P.-P., S.T.d.F., M.V.J.; resources M.V.J.; data curation: F.P., G.S.Z., M.P.-P.; writing—original draft preparation F.P., S.T.d.F., M.V.J.; writing—review and editing: F.P. and M.V.J.; visualization F.P., G.S.Z., M.P.-P.; supervision F.P., S.T.d.F., M.V.J.; project administration: F.P. and M.V.J.; funding acquisition M.V.J. All authors have read and agreed to the published version of the manuscript.

Funding: We thank FAPERJ (CNE E-26/202.780/2018) and CNPq (PDE 205072/2018-6) for funding FP. GSZ and MPP are doctoral students from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM), and receive doctoral fellowships from CONACYT, numbers 737920 and 694877, respectively. MVJ was funded by Dirección General de Asuntos del Personal Académico (DGAPA), Universidad Nacional Autónoma de México, UNAM (PAPIIT-IN201019).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Noller, H.F.; Woese, C.R. Secondary structure of 16S ribosomal RNA. *Science* **1981**, *212*, 403–411. [[CrossRef](#)]
2. Noller, H.F.; Hoffarth, V.; Zimniak, L. Unusual resistance of peptidyl transferase to protein extraction procedures. *Science* **1992**, *256*, 1416–1419. [[CrossRef](#)]
3. Lohse, P.A.; Szostak, J.W. Ribozyme-catalysed amino-acid transfer reactions. *Nature* **1996**, *381*, 442–444. [[CrossRef](#)] [[PubMed](#)]
4. Lanier, K.A.; Petrov, A.S.; Williams, L.D. The Central Symbiosis of Molecular Biology: Molecules in Mutualism. *J. Mol. Evol.* **2017**, *85*, 8–13. [[CrossRef](#)] [[PubMed](#)]
5. Vitas, M.; Dobovišek, A. In the Beginning was a Mutualism—On the Origin of Translation. *Orig. Life Evol. Biosph.* **2018**, *48*, 223–243. [[CrossRef](#)] [[PubMed](#)]
6. De Farias, S.T.; Prosdocimi, F. *A Emergência dos Sistemas Biológicos: Uma Visão Molecular da Origem da Vida*, 1st ed.; ArtecComCiencia: Rio de Janeiro, Brazil, 2019; ISBN 978-65-900624-1-3.
7. Prosdocimi, F.; José, M.V.; de Farias, S.T. The First Universal Common Ancestor (FUCA) as the Earliest Ancestor of LUCA's (Last UCA) Lineage. In *Evolution, Origin of Life, Concepts and Methods*; Pontarotti, P., Ed.; Springer: Berlin/Heidelberg, Germany, 2019; pp. 43–54. ISBN 978-3-030-30363-1.
8. Bokov, K.; Steinberg, S.V. A hierarchical model for evolution of 23S ribosomal RNA. *Nature* **2009**, *457*, 977–980. [[CrossRef](#)]

9. Petrov, A.S.; Bernier, C.R.; Hsiao, C.; Norris, A.M.; Kovacs, N.A.; Waterbury, C.C.; Stepanov, V.G.; Harvey, S.C.; Fox, G.E.; Wartell, R.M.; et al. Evolution of the ribosome at atomic resolution. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 10251–10256. [[CrossRef](#)]
10. Caetano-Anollés, G. Ancestral Insertions and Expansions of rRNA do not Support an Origin of the Ribosome in Its Peptidyl Transferase Center. *J. Mol. Evol.* **2015**, *80*, 162–165. [[CrossRef](#)]
11. Petrov, A.S.; Williams, L.D. The ancient heart of the ribosomal large subunit: A response to Caetano-Anollés. *J. Mol. Evol.* **2015**, *80*, 166–170. [[CrossRef](#)]
12. Rodnina, M.V. The ribosome as a versatile catalyst: Reactions at the peptidyl transferase center. *Curr. Opin. Struct. Biol.* **2013**, *23*, 595–602. [[CrossRef](#)]
13. Caetano-Anollés, G.; Caetano-Anollés, D. Computing the origin and evolution of the ribosome from its structure—Uncovering processes of macromolecular accretion benefiting synthetic biology. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 427–447. [[CrossRef](#)] [[PubMed](#)]
14. Lehmann, J. Induced fit of the peptidyl-transferase center of the ribosome and conformational freedom of the esterified amino acids. *RNA* **2016**, *23*, 229–239. [[CrossRef](#)] [[PubMed](#)]
15. Agmon, I.; Bashan, A.; Zarivach, R.; Yonath, A. Symmetry at the active site of the ribosome: Structural and functional implications. *Biol. Chem.* **2005**, *386*, 833–844. [[CrossRef](#)] [[PubMed](#)]
16. Belousoff, M.J.; Davidovich, C.; Zimmerman, E.; Caspi, Y.; Wekselman, I.; Rozenszajn, L.; Shapira, T.; Sade-Falk, O.; Taha, L.; Bashan, A.; et al. Ancient machinery embedded in the contemporary ribosome. *Biochem. Soc. Trans.* **2010**, *38*, 422–427. [[CrossRef](#)]
17. Tamura, K. Ribosome evolution: Emergence of peptide synthesis machinery. *J. Biosci.* **2011**, *36*, 921–928. [[CrossRef](#)]
18. Caetano-Anollés, G.; Sun, F.-J. The natural history of transfer RNA and its interactions with the ribosome. *Front. Genet.* **2014**, *5*, 5.
19. De Farias, S.T.; Rêgo, T.G.; José, M.V. Origin and evolution of the Peptidyl Transferase Center from proto-tRNAs. *FEBS Open Bio* **2014**, *4*, 175–178. [[CrossRef](#)]
20. Root-Bernstein, M.; Root-Bernstein, R. The ribosome as a missing link in the evolution of life. *J. Theor. Biol.* **2015**, *367*, 130–158. [[CrossRef](#)]
21. De Farias, S.T.; Rêgo, T.G.; José, M.V. Peptidyl Transferase Center and the Emergence of the Translation System. *Life* **2017**, *7*, 21. [[CrossRef](#)]
22. Demongeot, J.; Seligmann, H. Evolution of tRNA into rRNA secondary structures. *Gene Rep.* **2019**, *17*, 100483. [[CrossRef](#)]
23. Sayers, E.W.;avanaugh, M.; Clark, K.; Ostell, J.; Pruitt, K.D.; Karsch-Mizrachi, I. GenBank. *Nucleic Acids Res.* **2019**, *48*, D84–D86.
24. Needleman, S.B.; Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **1970**, *48*, 443–453. [[CrossRef](#)]
25. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276–277. [[CrossRef](#)]
26. Thompson, J.D.; Gibson, T.J.; Higgins, D.G. Multiple Sequence Alignment Using ClustalW and ClustalX. *Curr. Protoc. Bioinform.* **2002**, *00*, 2.3.1–2.3.22. [[CrossRef](#)] [[PubMed](#)]
27. Spang, A.; Stairs, C.W.; Dombrowski, N.; Eme, L.; Lombard, J.; Caceres, E.F.; Greening, C.; Baker, B.J.; Ettema, T.J.G. Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. *Nat. Microbiol.* **2019**, *4*, 1138–1148. [[CrossRef](#)] [[PubMed](#)]
28. Spang, A.; Saw, J.H.; Jørgensen, S.L.; Zaremba-Niedzwiedzka, K.; Martijn, J.; Lind, A.E.; Van Eijk, R.; Schleper, C.; Guy, L.; Ettema, T.J.G. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **2015**, *521*, 173–179. [[CrossRef](#)]
29. Crooks, G.E.; Hon, G.; Chandonia, J.-M.; Brenner, S.E. WebLogo: A Sequence Logo Generator. *Genome Res.* **2004**, *14*, 1188–1190. [[CrossRef](#)]
30. Zamudio, G.S.; José, M.V. Identity Elements of tRNA as Derived from Information Analysis. *Orig. Life Evol. Biosph.* **2017**, *48*, 73–81. [[CrossRef](#)]
31. Meilã, M. Comparing Clusterings by the Variation of Information. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, 24–27 August 2003: Proceedings*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 173–187, ISBN 978-3-540-40720-1.

32. Zamudio, G.S.; Palacios-Pérez, M.; José, M.V. Information theory unveils the evolution of tRNA identity elements in the three domains of life. *Theory Biosci.* **2020**, *139*, 77–85. [[CrossRef](#)]
33. Bernier, C.R.; Petrov, A.S.; Waterbury, C.C.; Jett, J.; Li, F.; Freil, L.E.; Xiong, X.; Wang, L.; Migliozi, B.; Hershkovits, E.; et al. RiboVision suite for visualization and analysis of ribosomes. *Faraday Discuss.* **2014**, *169*, 195–207. [[CrossRef](#)]
34. Reuter, J.S.; Mathews, D.H. RNAstructure: Software for RNA secondary structure prediction and analysis. *BMC Bioinform.* **2010**, *11*, 129. [[CrossRef](#)] [[PubMed](#)]
35. Kerpedjiev, P.; Hammer, S.; Hofacker, I.L. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics* **2015**, *31*, 3377–3379. [[CrossRef](#)] [[PubMed](#)]
36. Okonechnikov, K.; Golosova, O.; Fursov, M. Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics* **2012**, *28*, 1166–1167. [[CrossRef](#)] [[PubMed](#)]
37. Rother, M.; Rother, K.; Puton, T.; Bujnicki, J.M. ModeRNA: A tool for comparative modeling of RNA 3D structure. *Nucleic Acids Res.* **2011**, *39*, 4007–4022. [[CrossRef](#)]
38. Gong, S.; Zhang, C.; Zhang, Y. RNA-align: Quick and accurate alignment of RNA 3D structures based on size-independent TM-scoreRNA. *Bioinformatics* **2019**, *35*, 4459–4461. [[CrossRef](#)]
39. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem* **2004**, *25*, 1605–1612. [[CrossRef](#)]
40. Prosdocimi, F.; Linard, B.; Pontarotti, P.; Poch, O.; Thompson, J.D. Controversies in modern evolutionary biology: The imperative for error detection and quality control. *BMC Genom.* **2012**, *13*, 5. [[CrossRef](#)]
41. Schroeder, R.; Barta, A.; Semrad, K. Strategies for RNA folding and assembly. *Nat. Rev. Mol. Cell Biol.* **2004**, *5*, 908–919. [[CrossRef](#)]
42. Schultes, E.A.; Bartel, D.P. One sequence, two ribozymes: Implications for the emergence of new ribozyme folds. *Science* **2000**, *289*, 448–452. [[CrossRef](#)]
43. Brion, P.; Westhof, E. Hierarchy and dynamics of RNA folding. *Annu. Rev. Biophys. Biomol. Struct.* **1997**, *26*, 113–137. [[CrossRef](#)]
44. Draper, D.E. A guide to ions and RNA structure. *RNA* **2004**, *10*, 335–343. [[CrossRef](#)] [[PubMed](#)]
45. De Farias, S.T. Suggested phylogeny of tRNAs based on the construction of ancestral sequences. *J. Theor. Biol.* **2013**, *335*, 245–248. [[CrossRef](#)] [[PubMed](#)]
46. Jühling, F.; Mörl, M.; Hartmann, R.K.; Sprinzl, M.; Stadler, P.F.; Pütz, J. tRNAdb 2009: Compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.* **2009**, *37*, D159–D162. [[CrossRef](#)] [[PubMed](#)]
47. Korostelev, A.; Trakhanov, S.; Laurberg, M.; Noller, H.F. Crystal Structure of a 70S Ribosome-tRNA Complex Reveals Functional Interactions and Rearrangements. *Cell* **2006**, *126*, 1065–1077. [[CrossRef](#)]
48. Krupkin, M.; Matzov, D.; Tang, H.; Metz, M.; Kalaora, R.; Belousoff, M.J.; Zimmerman, E.; Bashan, A.; Yonath, A. A vestige of a prebiotic bonding machine is functioning within the contemporary ribosome. *Philos. Trans. R. Soc. B Biol. Sci.* **2011**, *366*, 2972–2978. [[CrossRef](#)]
49. De Farias, S.T.; José, M.V. Transfer RNA: The molecular demiurge in the origin of biological systems. *Prog. Biophys. Mol. Boil.* **2020**, *153*, 28–34. [[CrossRef](#)]
50. Prosdocimi, F.; de Farias, S.T. From FUCA To LUCA: A Theoretical Analysis on the Common Descent of Gene Families. *Acta Sci. Microbiol.* **2020**, *3*, 1–9.
51. Ivanov, V.I.; Bondarenko, S.A.; Zdobnov, E.M.; Beniaminov, A.D.; Minyat, E.E.; Ulyanov, N.B. A pseudoknot-compatible universal site is located in the large ribosomal RNA in the peptidyltransferase center. *FEBS Lett.* **1999**, *446*, 60–64. [[CrossRef](#)]
52. Polacek, N.; Mankin, A.S. The Ribosomal Peptidyl Transferase Center: Structure, Function, Evolution, Inhibition. *Crit. Rev. Biochem. Mol. Biol.* **2005**, *40*, 285–311. [[CrossRef](#)]
53. Chirkova, A.; Erlacher, M.D.; Clementi, N.; Żywicki, M.; Aigner, M.; Polacek, N. The role of the universally conserved A2450–C2063 base pair in the ribosomal peptidyl transferase center. *Nucleic Acids Res.* **2010**, *38*, 4844–4855. [[CrossRef](#)]
54. Davidovich, C.; Belousoff, M.J.; Wekselman, I.; Shapira, T.; Krupkin, M.; Zimmerman, E.; Bashan, A.; Yonath, A. The Proto-Ribosome: An ancient nano-machine for peptide bond formation. *Isr. J. Chem.* **2010**, *50*, 29–35. [[CrossRef](#)] [[PubMed](#)]
55. Terasaka, N.; Hayashi, G.; Katoh, T.; Suga, H. An orthogonal ribosome-tRNA pair via engineering of the peptidyl transferase center. *Nat. Chem. Biol.* **2014**, *10*, 555–557. [[CrossRef](#)] [[PubMed](#)]

56. Lehmann, J. Physico-chemical Constraints Connected with the Coding Properties of the Genetic System. *J. Theor. Biol.* **2000**, *202*, 129–144. [[CrossRef](#)] [[PubMed](#)]
57. Lehmann, J.; Cibils, M.; Libchaber, A. Emergence of a Code in the Polymerization of Amino Acids along RNA Templates. *PLoS ONE* **2009**, *4*, e5773. [[CrossRef](#)] [[PubMed](#)]
58. Zaucha, J.; Heddle, J.G. Resurrecting the Dead (Molecules). *Comput. Struct. Biotechnol. J.* **2017**, *15*, 351–358. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Copyright of Life (2075-1729) is the property of MDPI Publishing and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.

Apéndice 10

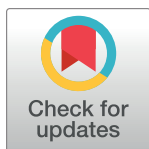
RESEARCH ARTICLE

Anticipation of ventricular tachyarrhythmias by a novel mathematical method: Further insights towards an early warning system in implantable cardioverter defibrillators

Gabriel S. Zamudio^{1*}, Manlio F. Márquez², Marco V. José^{1*}

1 Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, Ciudad de México, México, **2** Electrophysiology Department, Instituto Nacional de Cardiología Ignacio Chávez, Mexico City, Mexico

* gazaso92@gmail.com (GSZ); marcojose@biomedicas.unam.mx (MVJ)



OPEN ACCESS

Citation: Zamudio GS, Márquez MF, José MV (2020) Anticipation of ventricular tachyarrhythmias by a novel mathematical method: Further insights towards an early warning system in implantable cardioverter defibrillators. PLoS ONE 15(10): e0235101. <https://doi.org/10.1371/journal.pone.0235101>

Editor: Elena G. Tolkacheva, University of Minnesota, UNITED STATES

Received: June 8, 2020

Accepted: September 14, 2020

Published: October 1, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pone.0235101>

Copyright: © 2020 Zamudio et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The data underlying this study are third party. The data was downloaded from the Spontaneous Ventricular

Abstract

Implantable cardioverter defibrillators (ICD) are the most effective therapy to terminate malignant ventricular arrhythmias (VA) and therefore to prevent sudden cardiac death. Until today, there is no way to predict the onset of such VA. Our aim was to develop a mathematical model that could predict VA in a timely fashion. We analyzed the time series of R-R intervals from 3 groups. Two groups from the Spontaneous Ventricular Tachyarrhythmia Database (v 1.0) were analyzed from a set of 81 pairs of R-R interval time series records from patients, each pair containing one record before the VT episode (Dataset 1A) and one control record which was obtained during the follow up visit (Dataset 1B). A third data set was composed of the R-R interval time series of 54 subjects without a significant arrhythmia heart disease (Dataset 2). We developed a new method to transform a time series into a network for its analysis, the ϵ -regular graphs. This novel approach transforms a time series into a network which is sensitive to the quantitative properties of the time series, it has a single parameter (ϵ) to be adjusted, and it can trace long-range correlations. This procedure allows to use graph theory to extract the dynamics of any time series. The average of the difference between the VT and the control record graph degree of each patient, at each time window, reached a global minimum value of -2.12 followed by a drastic increase of the average graph until reaching a local maximum of 5.59 . The global minimum and the following local maxima occur at the windows 276 and 393, respectively. This change in the connectivity of the graphs distinguishes two distinct dynamics occurring during the VA, while the states in between the 276 and 393, determine a transitional state. We propose this change in the dynamic of the R-R intervals as a measurable and detectable “early warning” of the VT event, occurring an average of 514.625 seconds (8:30 minutes) before the onset of the VT episode. It is feasible to detect retrospectively early warnings of the VA episode using their corresponding ϵ -regular graphs, with an average of 8:30 minutes before the ICD terminates the VA event.

Tachyarrhythmia Database Version 1.0 from Medtronic Inc. (available at doi.org/10.13026/C25K5D). A set of 81 pairs of R-R interval time series records from different patients was obtained (Group 1A), each pair contains one record before the VT episode and one control record (CR) which was obtained during the follow up visit (Group 1B). A third data set composed of the R-R interval time series of 54 subjects without a significant arrhythmia heart disease was obtained (available at doi.org/10.13026/C2NK5R). The authors did not have special access privileges.

Funding: Gabriel S. Zamudio is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and a fellowship recipient from Consejo Nacional de Ciencia y Tecnología (CONACYT) (number: 737920). Marco V. José was financially supported by PAPIIT-IN201019, UNAM, México.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Implantable cardioverter defibrillators (ICD) are the cornerstone of sudden cardiac death prevention through termination of ventricular tachycardia/ventricular fibrillation. Although ICD shocks usually occur when the subject is unconscious, it could be very useful to patients and close relatives to have the possibility to know in advance, either seconds or minutes, when those malignant arrhythmias could occur in order to take appropriate preventive measures. We hypothesized that a novel mathematical analysis, ε -regular graphs, could perform such task.

Network theory possesses the capacity to abstractly represent interactions of any kind of entities. Currently, complex networks have arisen as a common way to tackle intricate dynamics [1]. A broad range of applications in different biological and medical areas abound. In the area of biology, they have been used to analyze a population's structure [2, 3], and pandemics [4, 5]. The use of protein-protein interaction networks coupled with information theory have led to discover potential therapeutic biomarkers on cancer research [6]. Integrative approaches for anticipating critical transitions have been proposed [7] in several phenomena, although the area of cardiology has not yet been explored. The terms “early warnings” and “tipping points” are still not part of the cardiologist community. Several methods have been developed to transform time series into networks for its analysis. Such methods include the visibility graphs method [8], and a plethora of its modifications [9, 10], which consider the topological properties of the time series, the recurrence analysis of time series [11, 12], and the analysis based on the phase space [13]. In this work, we usher in a new method to transform a time series into a network for its analysis, the ε -regular graphs. This novel approach transforms a time series into a network which is sensitive to the quantitative properties of the time series, it has a single parameter (ε) to be adjusted, and it can capture long-range correlations. This procedure permits using graph theory to extract the dynamics of any time series. As a direct application of ε -regular graphs, data from patients diagnosed with imminent ventricular tachyarrhythmias (VT) was analyzed. The heart activity is driven by the action of the opposing forces of the sympathetic and the parasympathetic nervous systems [14, 15]. The failure in heart function is the result of malfunctions in the myocardium, heart valves, pericardium, or the endocardium [16].

Limitations

The method proposed in this work requires further testing with patients whose clinical history is well documented and controlled, coupled with respiratory data. A significant clinical limitation of this work is the fact that this approach is restricted to patients with normal sinus rhythm and is unlikely to work in patients with atrial fibrillation or those with a pacemaker. In the former group because of the large variability of R-R intervals, and in the later because of fixed pacing rhythms.

Methods

A mathematical method to transform time series to networks

The method consists in assigning to each point in a time series a vertex in the network. Then, for a fixed value of the parameter $\varepsilon = \varepsilon_0$, any two points of the time series p_1, p_2 will be joined in the network, if and only if $|p_1 - p_2| \leq \varepsilon_0$; this means that two point of the time series will be adjacent in the ε -regular graph if the values of the points have a maximum difference of ε_0 . For illustrative purposes, we show a diagram of the algorithm in Fig 1. In Fig 1A, we show a time series of ten points; the values of the points are p_1, p_7 , and $p_{10} = 0.2$; p_2 and p_4 are equal to 0.29; $p_5 = p_8 = 0.38$; $p_3 = p_6 = p_9 = 0.7$. The ε -regular graph is constructed with a parameter

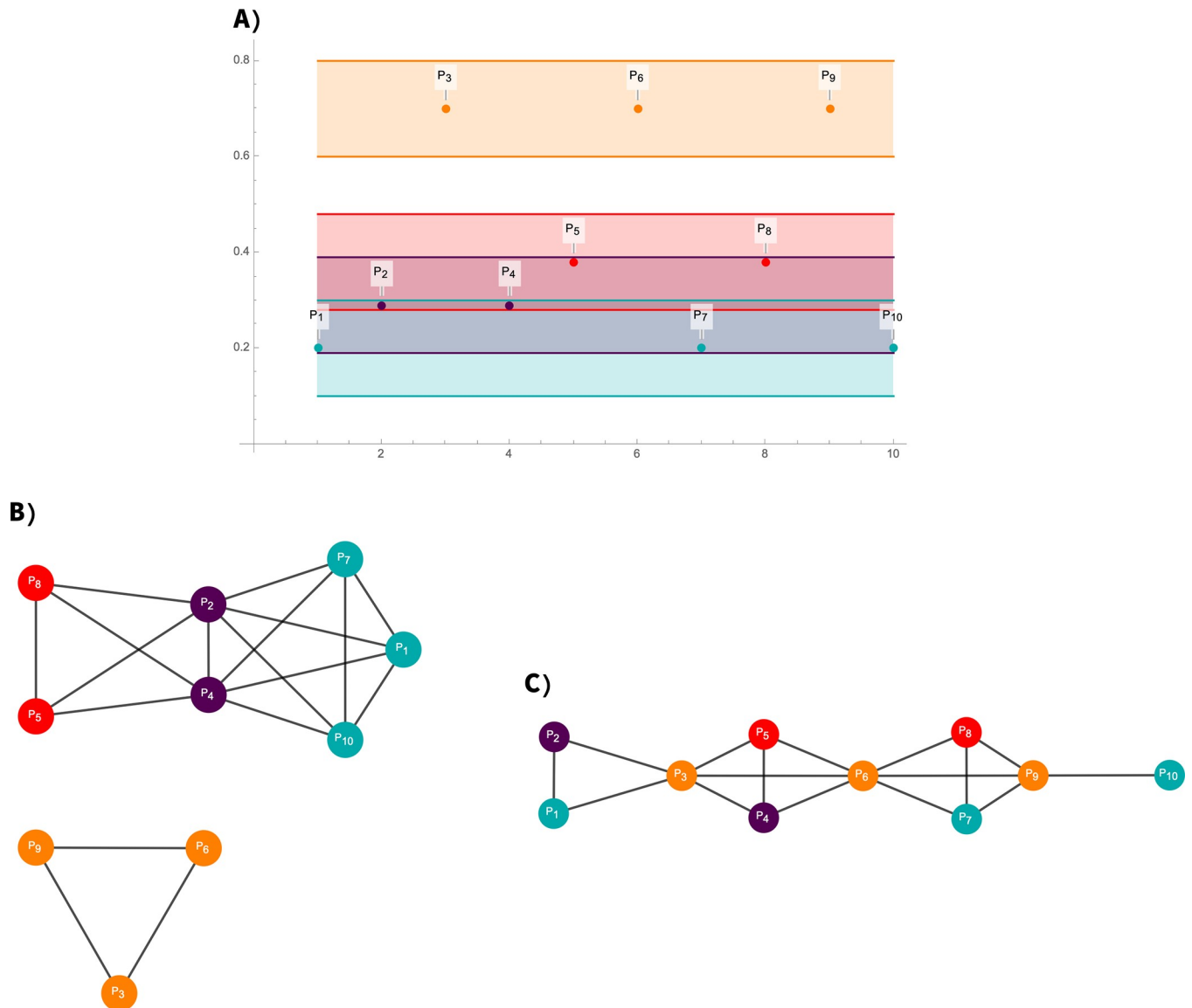


Fig 1. Diagram of the ϵ -graph algorithm. In (A), a time series of ten points from which an ϵ -regular graph is constructed with a parameter value of $\epsilon_0 = 0.1$. In (B), intervals of width $\epsilon_0 = 0.1$ are drawn around each point of the time series. For a given point p of the time series, all the point lying inside the interval of width ϵ_0 will be adjacent to p in the corresponding ϵ -regular graph. Note that points with a higher value of 0.7, belong to a different component than the rest of time series reflecting its outlier nature. Also note that the periodic values, p_3, p_6 and p_9 form a subgraph. Points with the same value that are not periodic will form a complete subgraph, such as the points p_1, p_7 and p_{10} . In (C), the time series is transformed to a graph using the visibility-graph algorithm.

<https://doi.org/10.1371/journal.pone.0235101.g001>

value of $\epsilon_0 = 0.1$. In Fig 1B, intervals of width $\epsilon_0 = 0.1$ are drawn around each point of the time series. For a given point p of the time series, all the point lying inside the interval of width ϵ_0 will be adjacent to p in the corresponding ϵ -regular graph.

Other algorithms to convert time series into graphs have been developed but with qualitative instead of quantitative rules for determining the adjacencies in the corresponding graphs (visibility plots). The algorithm for the visibility graphs determines the adjacencies by analyzing the lines joining the points of the time series [8] as observed in Fig 1C. The visibility graph algorithm confers to its graph properties that strongly differ to our ϵ -regular graph derived from the same time series because the former does not capture the quantitative properties of a

time series as the ϵ -regular graph do. When considering the outliers of a time series, in a visibility graph, the extremely high or low values of a time series would be “visible” from almost all the rest of the points of the time series, and thus, according to the visibility graph algorithm would be highly connected and act as a hub. This would remain true even if the outlier value were not so extremely contrasting. In contrast, an outlier in a ϵ -regular graph would have different behaviors according to its value. If the outlier value is significantly different, it will be assigned to a different component in the ϵ -regular graph: If the suspected outlier’s point value is not so different to the rest of the time series, it will remain in the same component. This behavior is reflected in the time series of Fig 1, where points with a higher value of 0.7, belong to a different component than the rest of time series reflecting its outlier nature. If instead of the value 0.7, the values were set to 0.48, they will still be higher than the rest of the time series but will remain in the same graph component of the rest of the time series points since the ϵ_0 value is set at $\epsilon_0 = 0.1$, and the points with the second-highest value are equal to 0.38.

From the definition of adjacencies in an ϵ -regular graph, it is directly derived that on a given set of points, in which the points of the set have a value difference up to ϵ_0 among them, they will be joined and thus they will form a complete subgraph. This result is useful when considering time series with regular or periodic values. The periodic values will form a complete subgraph, as the points p_3, p_6 , and p_9 in Fig 1B. Also, points with the same value that are not periodic will form a complete subgraph, such as the points p_1, p_7 , and p_{10} (Fig 1) that form the complete graph K_3 . A similar property is not inherited in visibility graphs. In visibility graphs, periodic or regular points from the time series might in some cases not be adjacent, as some points might block the visibility condition for them to be adjacent. What can be rescued from visibility graphs is the short-term correlations of the time series.

As proof of concept, two time series were simulated from theoretical frameworks and their corresponding ϵ -regular graphs were derived. Time series of 10^4 points were simulated, the first time series was obtained from a standard normal probability distribution, and the second from a standard Brownian motion valued at integer times. The value of the parameter ϵ_0 was set as the standard deviation, and half of the standard deviation of both the normal distribution and the Brownian motion, which correspond to 1 and 1/2 respectively, in both cases. The degree centrality measure was calculated for the resulting graphs (Fig 2). The values of the centrality measure were standardized to the unit interval. The degree distributions for the graphs constructed from the Brownian motion approximates a Gaussian curve despite its multiple modes; and the distribution from points sampled from a normal distribution approximates a lognormal distribution. The distributions of the degree measurements maintain, up to some extent, the statistical properties from the time series they were derived from. The difference between any two values of the Brownian motion follows a normal distribution, and this property is shown by the ϵ -regular graphs in the degree centrality measure when setting the ϵ -parameter equal to values related to the standard deviation.

The computer program is found in Supporting Information I.

Application to imminent VT

As a direct application of the ϵ -regular graphs algorithm, a set of time series related to VT was analyzed. The data was downloaded from the Spontaneous Ventricular Tachyarrhythmia Database Version 1.0 from Medtronic Inc. (available at <http://physionet.org/physiobank/database/mvtdb/>) [17]. A set of 81 pairs of R-R interval time series records from different patients was obtained (Group 1A), each pair contains one record before the VT episode and one control record (CR) which was obtained during the follow up visit (Group 1B). A third data set composed of the R-R interval time series of 54 subjects without a significant

Degree centrality of the applied ε -regular graph algorithm

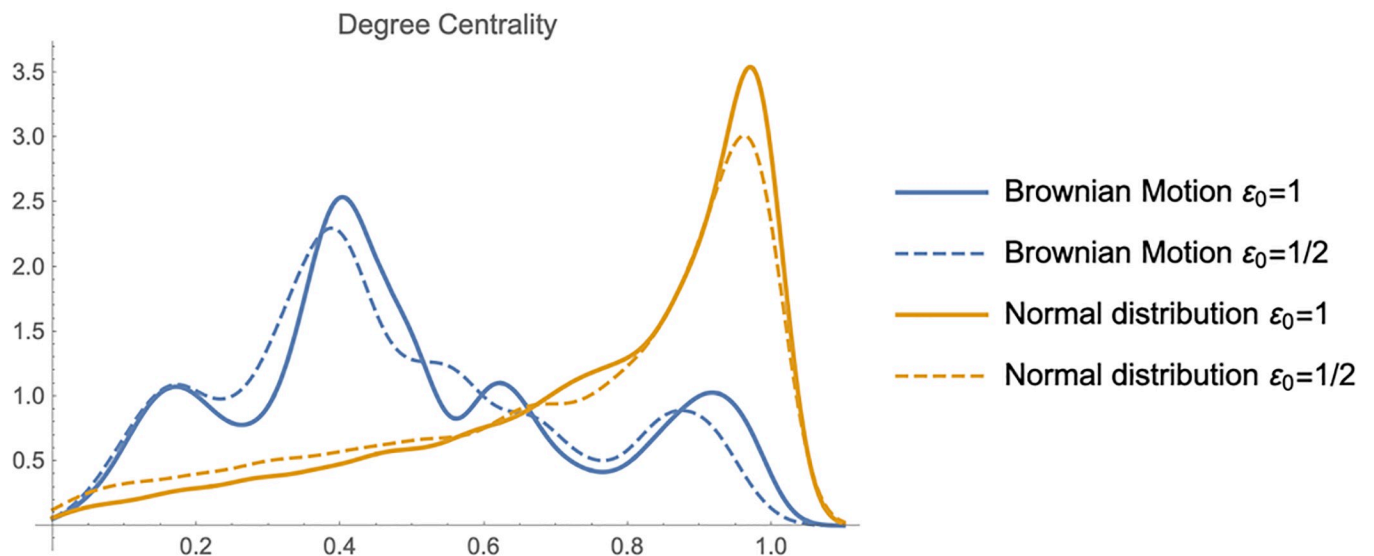


Fig 2. Degree centrality of the applied ε -graph algorithm. Two time series were simulated from theoretical frameworks and their corresponding ε -regular graphs were derived. Time series of 10,000 points were simulated, the first time series was obtained from a standard normal probability distribution (solid and dashed yellow curves), and the second from a standard Brownian motion (solid and dashed blue curves) valued at integer times. The value of the parameter ε_0 was set as the standard deviation, and half of the standard deviation of both the normal distribution and the Brownian motion, which correspond to 1 and 1/2, respectively, in both cases. The degree distributions for the graphs constructed from the Brownian motion approximates a Gaussian curve despite its multiple modes; and the distribution from points sampled from a normal distribution approximates a lognormal distribution.

<https://doi.org/10.1371/journal.pone.0235101.g002>

arrhythmia heart disease was obtained (available at: <https://physionet.org/physiobank/database/nsrdb/>) [16]. This third dataset of time series will be hereafter considered as healthy subjects (HS) and denoted as Group 2.

The time series for groups 1A and 1B were cropped to the same length starting from the end to make them directly comparable (985 points), the start of the VT episode is at the last point of the time series. The Group 2 (HS) series were also cropped by subsampling a random set of sequential points of each subject that match the length of the VT and CR time series. The VT, CR, and HS time series of each patient were subdivided in time series of 60 points with a sliding window method with an offset of one point. This would allow analyzing the change of the series in time to detect an early warning of the VT episode. The time series on each window were transformed into graphs with the ε -regular graph algorithm setting the parameter value at $\varepsilon_0 = 0.04$.

The average degree of the graphs from each window of each subject was calculated and averaged among the subjects, for the three different time series. This process results in a time series of the average degree of the graphs representing the three different states of the subjects (Fig 3). The datasets VT and CR arise from the same subjects, so a direct comparison of subjects prior to a VT episode and in a normal stage is possible. The average of the difference between the VT and the CR graph degree of each patient, at each time window (Fig 3), reaches a global minimum value of -2.12 , followed by a drastic increase of the average graph until reaching a local maximum of 5.59 . The global minimum and the following local maxima occur at the windows 276 and 393, respectively. This change in the connectivity of the graphs distinguishes two distinct dynamics occurring during the ventricular tachyarrhythmia, while the states in between the 276 and 393, determine a transitional state. We propose this change in

The ε -regular graph algorithm and the detrended fluctuation analysis

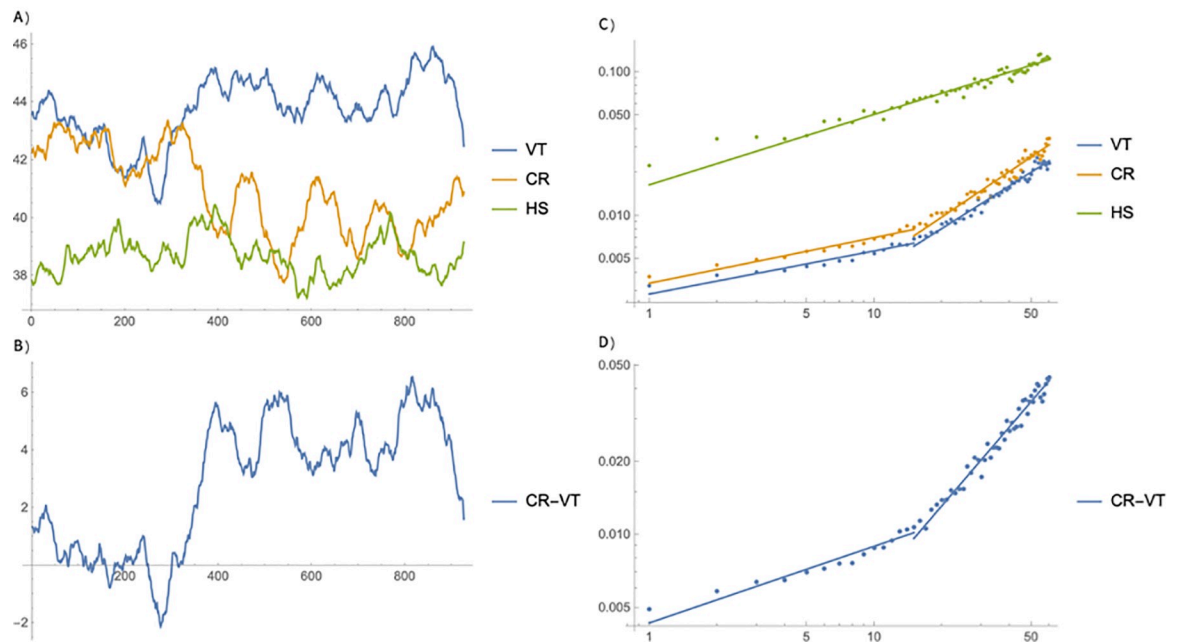


Fig 3. Early warning of a VT event. Comparison of the ε -graph algorithm and detrended fluctuation analysis. In (A), a set of 81 pairs of RR interval time series records from different patients was obtained (Group 1A), each pair contains one record before the VT episode and one control record (CR yellow solid curve) which was obtained during the follow up visit (Group 1B). A third data set composed of the RR interval time series of 54 healthy subjects (HS green solid curve) without a significant arrhythmia heart disease was obtained, denoted also as Group 2. In (B), the result of applying the detrended fluctuation analysis (DFA) to the RR series displayed in (A), are shown. The start of the VT episode is at the last point of the time series. The time series on each window were transformed into ε -graphs setting the parameter value at $\varepsilon_0 = 0.04$. The average degree of the graphs from each window of each subject was calculated and averaged among the subjects, for the three different time series. This process results in a time series of the average degree of the graphs representing the three different states of the subjects. In (C), the average of the difference between the VT and the CR graph degree of each patient, at each time window reaches a global minimum value of -2.12, followed by a drastic increase of the average graph until reaching a local maximum of 5.59. The global minimum and the following local maxima occur at the windows 276 and 393, respectively. This change in the connectivity of the graphs distinguishes two distinct dynamics occurring during the ventricular tachyarrhythmia, while the states in between the 276 and 393, determine a transitional state. In (D), the corresponding DFA of the curve obtained in (C) is shown.

<https://doi.org/10.1371/journal.pone.0235101.g003>

the dynamic of the R-R intervals as a measurable and detectable early warning of the VT event, occurring an average of 514.625 seconds (8:30 minutes) before the start of the VT episode. The count of 514.625 seconds corresponds to the sum of the average of the time lapses of the last R-R intervals starting from the point 276 to when the VT episode begins at point 985.

The optimization of the parameter ε is based on the statistical parameters of the R-R time series. The average, minimum and maximum distance between two points of the R-R time series are: 0.043, 1.11×10^{-16} , 0.12, respectively. Thus, the ε value of 0.04 approximates the mean of the differences. In Supporting Information II, we show the degree time series when the values 0.01, 0.02, 0.03, 0.04, and 0.06 are assigned to the parameter ε (S1 Fig). Note that the overall pattern is preserved. In particular, the abrupt change of the dynamics is captured by the different values of the parameter ε .

Comparison with detrending fluctuation analysis

A widely used procedure used in the analysis of data originated from diverse heart records is the application of the Detrended fluctuation analysis (DFA) method. DFA is a mathematical

linear method to analyze time series by removing the linear trend of time series divided into smaller windows. This method is of special use to address nonstationary time series [18]. Results from the DFA method are commonly graphed in a log-log scale and the scaling exponent of the time series is estimated from a least-square fit of a linear model. The scaling exponent measures the correlation in the noise and approximates the Hurst exponent of the time series.

The DFA method was applied to the mean time series of the 3 groups (Fig 3A), and the difference between the mean VT and CR subjects (Fig 3B). From the DFA of the different states of the subjects it can be observed that the time series from the HS subjects possess the same scaling properties at short- and long-time lengths, which is deduced from the fact that the DFA approximates a linear model. On the other hand, the DFA analysis from the VT and CR subjects show that their time series possess two different scales of autocorrelation, which is related to the two different linear models fitting the DFA of VT and CR. The VT and CR time series behave similarly in the sense that both exhibit different scaling properties for short correlations (windows with 15 or less points) and a different scaling factor for long correlations (15 or more points). Since the slope of the linear models fitted to VT, 0.24 and 0.95 for short and long correlations, respectively, and CR, 0.25 for short and 1.01 for long correlations (Fig 3C), are practically the same, then, their corresponding Hurst exponents will be the same, which results in that the VT and CR behave similarly regardless if short or long correlations are assessed. This is validated by the fact that the slopes of the two linear models fitted to the DFA of VT-CR, 0.31 and 1.08 for short and long correlations, approximates the ones obtained when analyzing VT and CR separately (Fig 3D). The DFA method is capable to discern the different scaling properties occurring on the Groups 1A (VT) and 1B (CR) patients as compared to the Group 2 (HS) subjects. However, the DFA results are not varying in time, and hence this method is not capable of discerning an early warning for VT.

Discussion

In this work we propose a novel parametric method to analyze time series by transforming them into networks. By using this method, it is possible to apply the graph and network theory in the analysis of time series. Herein, a direct application of the ϵ -regular graph method is herein shown by using time series data derived from patients with ventricular heart tachyarrhythmia disease. The application of the ϵ -regular graph method, using a sliding window framework, detected a potential early warning of the disease that it is not detectable using the current linear methods available for the analysis of time series. The ϵ -regular graphs differ from the visibility graph method as the former is a parametric quantitative method and the latter is a qualitative approach. The adjustable parameter ϵ in ϵ -regular graphs, determines the sensitivity of the transformation of time series to networks. By varying the parameter, it is possible to obtain a range of graphs going from graphs in which a vertex is only connected to other ones having the same value, up to completely connected graphs. An inverse transformation, from a network to a time series, would be possible if there exists a compendium of graphs derived from the same time series using different ϵ values. Then, if needed, the original time series can be inferred using the different adjacencies from the ϵ -regular graphs. An inverse transformation that faithfully recovers the time series is not possible for visibility graphs. Since visibility graphs is a qualitative methodology, the values of a time series derived from these graphs would vary in an interval, whose length would be different for each point in the time series. The framework of complex networks for analyzing heart rate variability data towards the detection of early warnings and the design of clinical tools for disease management has been considered before as other nonlinear methods [19]. Visibility graphs have been applied to

the analysis of congestive heart failure [20]. In here, a statistical analysis of the scale-freeness of the obtained network is used for the detection of early stages of the disease. In a broader analysis, several summary statistics of a horizontal visibility network have been proposed as useful for the analysis of heart rate variability [21]. In general, the use of summary statistics for the detection of early warnings in a transition of dynamical state may be difficult since such statistics may rely on inadequate data or other factors [22]. Other studies have shown that the incorporation of respiration signals to the electrocardiogram data increase the detection of a VT episode [23]. Hitherto, the effect of the vagus nerve in the heart activity has been recently investigated [25]. Different techniques based on other methodologies and data have shown different times before the VT episode occurs [23, 24]. Any predictor, regardless of the methodology must clearly distinguish a VT episode from the usual cardiac arrhythmias of each patient to avoid false positive detection. So far, the low heart rate variability has been considered as the single predictor of heart failure, although the forces for the acceleration and deceleration in heart activity have been shown to be uncoupled [25]. The device used by the patient has high impact on any method for the detection of early warnings of a cardiac malfunction, as it has been shown that ICDs can detect QT variability in near-field or far-field right ventricular intracardiac electrogram [26]. ICD are excellent machines devoted to terminating VT and they have proved its efficacy to prevent sudden cardiac death in different clinical settings. The performance of the algorithms has been tested first to detect the VT and to provide appropriate shocks. Then, algorithms were improved to avoid unnecessary (“inappropriate”) discharges to the patient. In recent years there has been a small but strong movement in the medical community towards the possibility of alerting the patient when an ICD shock is going to occur. This possibility is not minor. From a clinical point of view, such alert could permit the patient or his close relatives to take appropriate measures before the shock takes place. A new window of opportunity (clinical interventions) could be generated if a software could be able to detect with some seconds or, even better, minutes, the possibility of an imminent ICD shock. Until today, there is no such possibility. The present retrospective study sheds light of a possible mathematical analysis that could detect “early warnings” of an appropriate ICD shock for VT with an average of 8:30 minutes. The process of optimization of the ε parameter value requires a more extensive clinical experimentation. It stands to reason that the parameter value is specific for each individual patient and it ought to be tuned from the complete set of clinical parameters of the patient to avoid false-positives and false-negatives. In a more general case, it is also probable that the parameter value of ε is not fixed throughout the day and is dependent of the circadian rhythm of the patient.

Obviously, this mathematical application should be tested prospectively, but this can only be done if implemented into the software of the ICD. Collaboration with ICD industry is vital to achieve such goal.

Conclusions

Early warnings of the VA episode could be detected using their corresponding ε -regular graphs, even 8:30 minutes before the ICD comes into action. A prospective study is warranted to further corroborate this finding.

Supporting information

S1 File. The script was developed using Wolfram Mathematica 12.0.

(NB)

S1 Fig. The time series of the R-R intervals for the VT episode are subtracted from the control group from an average of the records of 81 patients. The regular graphs are derived considering ε values of 0.01, 0.02, 0.03, 0.04, 0.05, 0.06 and a window of 60 points. The degree centrality is averaged for each window and plotted in this figure. (TIFF)

Acknowledgments

We thank Juan R. Bobadilla for technical computer support.

Author Contributions

Conceptualization: Gabriel S. Zamudio, Marco V. José.

Data curation: Gabriel S. Zamudio.

Formal analysis: Gabriel S. Zamudio, Marco V. José.

Funding acquisition: Marco V. José.

Investigation: Gabriel S. Zamudio, Manlio F. Márquez, Marco V. José.

Methodology: Gabriel S. Zamudio, Marco V. José.

Project administration: Marco V. José.

Resources: Marco V. José.

Software: Gabriel S. Zamudio.

Supervision: Manlio F. Márquez, Marco V. José.

Validation: Gabriel S. Zamudio, Manlio F. Márquez.

Visualization: Manlio F. Márquez.

Writing – original draft: Gabriel S. Zamudio, Marco V. José.

Writing – review & editing: Gabriel S. Zamudio, Manlio F. Márquez, Marco V. José.

References

1. Albert R.; Barabási A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* 2002, 74, 47–97.
2. Dyer R.J.; Nason J.D.; Garrick R.C. Landscape modelling of gene flow: Improved power using conditional genetic distance derived from the topology of population networks. *Mol. Ecol.* 2010, 19, 3746–3759. <https://doi.org/10.1111/j.1365-294X.2010.04748.x> PMID: 20723052
3. Dyer R.J.; Nason J.D. Population Graphs: the graph theoretic shape of genetic structure. *Mol. Ecol.* 2004, 13, 1713–1727. <https://doi.org/10.1111/j.1365-294X.2004.02177.x> PMID: 15189198
4. Simini F.; González M.C.; Maritan A.; Barabási A.L. A universal model for mobility and migration patterns. *Nature* 2012, 484, 96–100. <https://doi.org/10.1038/nature10856> PMID: 22367540
5. Barrio R.A.; Varea C.; Govezensky T.; Jose M. V. Modeling the geographical spread of influenza A (H1N1): the case of Mexico. *Appl. Math. Sci.* 2016, 7, 2143–2176.
6. Juarez-Flores A.; José M. V. Multivariate Entropy Characterizes the Gene Expression and Protein-Protein Networks in Four Types of Cancer. *Entropy* 2018, 20, 154.
7. Scheffer M.; Bascompte J.; Brock W. a; Brovkin V; Carpenter S.R; Dakos V; et al. Early-warning signals for critical transitions. *Nature* 2009, 461, 53–59. <https://doi.org/10.1038/nature08227> PMID: 19727193
8. Lacasa L.; Luque B.; Ballesteros F.; Luque J.; Nuño J.C. From time series to complex networks: The visibility graph. *Proc. Natl. Acad. Sci.* 2008, 105, 4972–4975. <https://doi.org/10.1073/pnas.0709247105> PMID: 18362361
9. Gonçalves B.A.; Carpi L.; Rosso O.A.; Ravetti M.G. Time series characterization via horizontal visibility graph and Information Theory. *Phys. A Stat. Mech. its Appl.* 2016, 464, 93–102.

10. Bezsudnov I. V.; Snarskii A.A. From the time series to the complex networks: The parametric natural visibility graph. *Phys. A Stat. Mech. its Appl.* 2014, 414, 53–60.
11. Marwan N.; Donges J.F.; Zou Y.; Donner R. V.; Kurths J. Complex network approach for recurrence analysis of time series. *Phys. Lett. Sect. A Gen. At. Solid State Phys.* 2009, 373, 4246–4254.
12. González H.; Infante O.; Perez-Grovas H.; José M.V.; Lerma C. Nonlinear dynamics of heart rate variability in response to orthostatism and hemodialysis in chronic renal failure patients: Recurrence analysis approach. *Medical Engineering and Physics* 2013, 35, 178–187. <https://doi.org/10.1016/j.medengphy.2012.04.013> PMID: 22647839
13. Wang M.; Tian L. From time series to complex networks: The phase space coarse graining. *Phys. A Stat. Mech. its Appl.* 2016, 461, 456–468.
14. Schwartz P.J. Vagal Stimulation for Heart Diseases: From Animals to Men. *Circ. J.* 2011, 75, 20–27. <https://doi.org/10.1253/circj.cj-10-1019> PMID: 21127379
15. Triposkiadis F.; Karayannis G.; Giamouzis G.; Skoularigis J.; Louridas G.; Butler J. The Sympathetic Nervous System in Heart Failure. Physiology, Pathophysiology, and Clinical Implications. *J. Am. Coll. Cardiol.* 2009, 54, 1747–1762. <https://doi.org/10.1016/j.jacc.2009.05.015> PMID: 19874988
16. Patel M.R.; White R.D.; Abbara S.; Bluemke D.A.; Herfkens R.J.; Picard M.; et al. 2013 ACCF/ACR/AHA/ASNC/SCCT/SCMR Appropriate Utilization of Cardiovascular Imaging in Heart Failure. *J. Am. Coll. Cardiol.* 2013, 61, 2207–2231. <https://doi.org/10.1016/j.jacc.2013.02.005> PMID: 23500216
17. Goldberger A.L.; Amaral L.A.N.; Glass L.; Hausdorff J.M.; Ivanov P.C.; Mark R.G.; et al. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation* 2000, 101. <https://doi.org/10.1161/01.cir.101.1.101> PMID: 10618311
18. Bryce R.M.; Sprague K.B. Revisiting detrended fluctuation analysis. *Sci. Rep.* 2012, 2, 315. <https://doi.org/10.1038/srep00315> PMID: 22419991
19. Henriques T.; Ribeiro M.; Teixeira A.; Castro L.; Antunes L.; Costa-Santos C. Nonlinear Methods Most Applied to Heart-Rate Time Series: A Review. *Entropy* 2020, 22, 309.
20. Bhaduri A.; Bhaduri S.; Ghosh D. Visibility graph analysis of heart rate time series and bio-marker of congestive heart failure. *Phys. A Stat. Mech. its Appl.* 2017, 482, 786–795.
21. Madl T. Network analysis of heart beat intervals using horizontal visibility graphs. In Proceedings of the Computing in Cardiology; 2016; Vol. 43, pp. 733–736.
22. Boettiger C.; Hastings A. Quantifying limits to detection of early warning for critical transitions. *J. R. Soc. Interface* 2012, 9, 2527–2539. <https://doi.org/10.1098/rsif.2012.0125> PMID: 22593100
23. Lee H.; Shin S.-Y.; Seo M.; Nam G.-B.; Joo S. Prediction of Ventricular Tachycardia One Hour before Occurrence Using Artificial Neural Networks. *Sci. Rep.* 2016, 6, 32390. <https://doi.org/10.1038/srep32390> PMID: 27561321
24. Au-Yeung W.-T.M.; Reinhall P.G.; Bardy G.H.; Brunton S.L. Development and validation of warning system of ventricular tachyarrhythmia in patients with heart failure with heart rate variability data. *PLoS One* 2018, 13, e0207215. <https://doi.org/10.1371/journal.pone.0207215> PMID: 30427880
25. Hu W.; Jin X.; Zhang P.; Yu Q.; Yin G.; Lu Y.; et al. Deceleration and acceleration capacities of heart rate associated with heart failure with high discriminating performance. *Sci. Rep.* 2016, 6, 23617. <https://doi.org/10.1038/srep23617> PMID: 27005970
26. Tereshchenko L.G.; Fetis B.J.; Domitrovich P.P.; Lindsay B.D.; Berger R.D. Prediction of Ventricular Tachyarrhythmias by Intracardiac Repolarization Variability Analysis. *Circ. Arrhythmia Electrophysiol.* 2009, 2, 276–284.

Copyright of PLoS ONE is the property of Public Library of Science and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.

Apéndice 11



OPEN

Novel gene signatures for stage classification of the squamous cell carcinoma of the lung

Angel Juarez-Flores, Gabriel S. Zamudio & Marco V. José✉

The squamous cell carcinoma of the lung (SCLC) is one of the most common types of lung cancer. As GLOBOCAN reported in 2018, lung cancer was the first cause of death and new cases by cancer worldwide. Typically, diagnosis is made in the later stages of the disease with few treatment options available. The goal of this work was to find some key components underlying each stage of the disease, to help in the classification of tumor samples, and to increase the available options for experimental assays and molecular targets that could be used in treatment development. We employed two approaches. The first was based in the classic method of differential gene expression analysis, network analysis, and a novel concept known as network gatekeepers. The second approach was using machine learning algorithms. From our combined approach, we identified two sets of genes that could function as a signature to identify each stage of the cancer pathology. We also arrived at a network of 55 nodes, which according to their biological functions, they can be regarded as drivers in this cancer. Although biological experiments are necessary for their validation, we proposed that all these genes could be used for cancer development treatments.

As GLOBOCAN reported in 2018, lung cancer was the first cause of deaths and new cases by cancer worldwide¹. Squamous cell carcinoma of the lung (SCC) is one type of lung cancer which comprises approximately 30% of all lung cancer cases. The available molecular targets for use in the treatment of SCC of the lung are behind of other types of cancer^{2–4}. Recent advances in the treatment have been achieved using immunotherapy as nivolumab and pembrolizumab and some clinical trials are being conducted to test molecular targets^{3,4}. Some efforts to understand the basis of the disease have been made using gene expression profiles, DNA sequencing and SNP arrays². However, there are few preclinical murine models, some SCC of the lung cell lines have errors in their classification and molecular targets usually found in other types of lung cancer as lung adenocarcinoma are rarely present in SCC of the lung⁴. Lung cancer is classified into two wide groups as follows: Small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). NSCLC represents 85% of all lung cancer cases. From this group the most prevalent are the adenocarcinoma and the squamous cell carcinoma of the lung^{2,5}. Lung cancer survival is less than 5% after 5 years and most of them metastasize. Most of the time lung cancer is detected in advanced stages in which treatment is less effective. The best treatment is surgery although the effectivity of the treatment is linked to early stages of the disease^{6–9}. Smoking is considered as a risk factor associated to lung cancer development². Network analysis is widely used in different areas including biological sciences with a wide variety of results. There are different metrics that can be obtained from networks as the hubs which are commonly referred as the most connected nodes which lead to network instability if they are perturbed^{10–13}. Besides, other network measures as betweenness and multivariate entropy have been used to analyze cancer networks to find putative potential targets for cancer disease^{14,15}. We previously identified a set of nodes which due to its biological and network properties we called them network gatekeepers¹⁶. The latter was done by visual inspection. Gatekeepers have few nearest-neighbor interactions with other proteins, but these proteins have plenty of interactions. Gatekeepers might not be detected by standard differential gene expression analyses.

In this work, we use clustering centrality as a metric for a better and quicker identification of gatekeepers¹⁶. Machine learning algorithms have been applied to a wide variety of phenomena¹⁷. Health sciences have a special interest in the applications of these techniques due to the vast data publicly available with the objective to achieve better diagnosis and treatments of diseases. Some of the analyzed data with these approaches include analysis of histopathological images^{18–20}. In this article, we make an analysis of the carcinogenic process of the squamous cell carcinoma of the lung using cutting-edge techniques as network and machine learning analyses to obtain sets of genes which could function as a signature to aid in the classification of patient tumor samples into one of the

Theoretical Biology Group, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, 04510 Ciudad Universitaria, Mexico. ✉email: marcojose@biomedicas.unam.mx

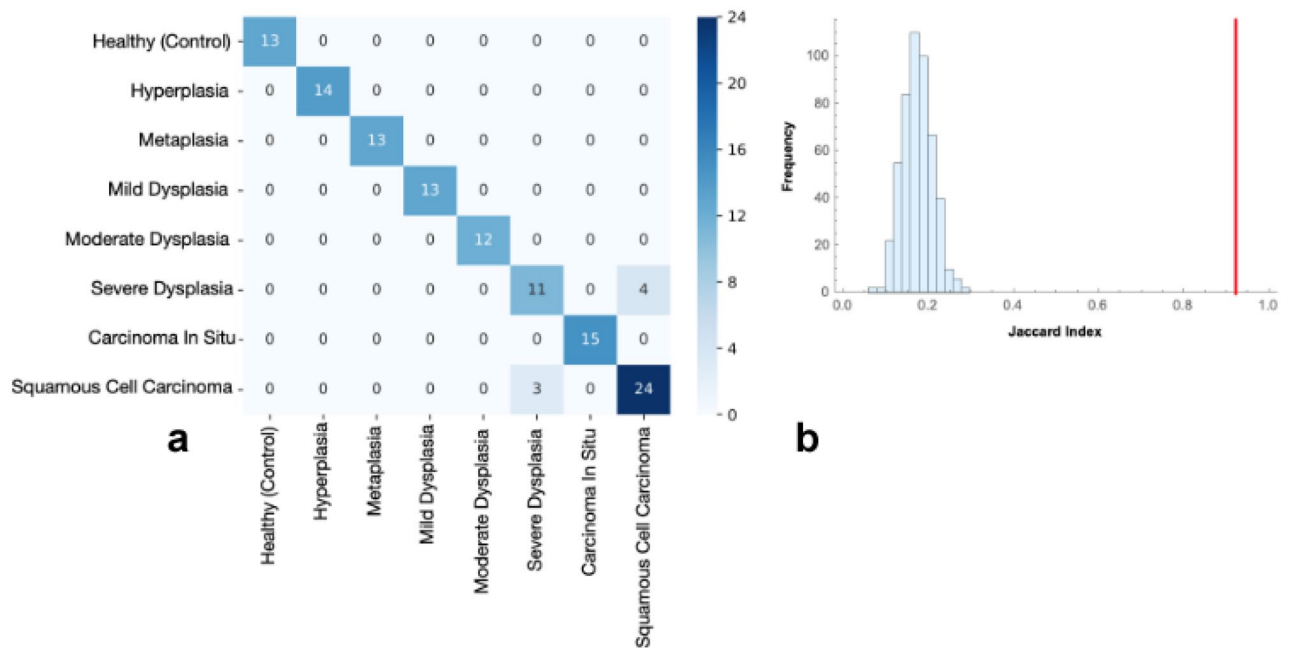


Figure 1. (a) Confusion matrix of the model trained with the 15 genes selected using the parameter reduction method. On the x-axis is the true classification and on the y-axis is the predicted classification for each of the 122 patient records. (b) Histogram of the Jaccard indexes from 500 trained models with random sets of 15 genes; In red the Jaccard index of the trained model with the 15 genes selected using the parameter reduction method. Figures were made using the library matplotlib of Python.

carcinogenic process stages and to increase the available options for experimental assays and molecular targets that could be used in treatment development. Although further biological experimental validation is needed.

Results

Carcinoma-stage classification model derived from machine learning. Data collected in GSE33479 was used to train a supervised machine learning algorithm. A logistic regression model was trained to classify the eight stages of the small cell lung carcinoma. Logistic regression models have been shown to provide accurate non-linear classification models of complex data²¹. A parameter reduction procedure was applied to the trained model. For this, the parameters of the model were standardized so that the parameters follow a standard normal distribution. The parameters whose value was beyond 0.78 from the mean were selected as relevant parameters. The procedure of model training and parameter selection was applied two times. On first parameter reduction, a total of 800 genes out of 41,067 were selected; on the second round a total of 15 relevant genes were selected. When using the subset of 800 genes a logistic regression model was trained and tested with records of all 122 patients records with all correctly classified, when considering the set of 15 genes the trained model was able to correctly classify the healthy stage and the first stages of small cell carcinoma and presented 7 cases of misclassification on later stages Fig. 1a. A neutral control was designed by considering a total of 500 random sets of 15 genes, on each random set a logistic regression model was trained, and its accuracy was measured using the Jaccard index and compared with the Jaccard index of a model trained with the selected set of 15 genes. The Jaccard index measures the proportion of correctly categorized cases by the model Fig. 1b. The Jaccard index from the set of genes derived from the parameter reduction method was 0.92 whereas for the random sets the maximum Jaccard index was 0.29. When considering the set of 15 selected genes coupled with the 26 genes identified from previous analysis on PPI networks resulted in a trained model with a Jaccard index of 1. APID PPI data was used for network analysis of the results from the implemented machine learning technique for the first glance results of approximately 800 genes. APID was used due to better coverage of most part of the 800 genes.

Differential gene expression analysis. The first step was to carry out an exploratory network analysis which is shown in Figs. 2a,b. These networks are obtained from joining the results from the Differential Gene Expression (DGE) to Mentha network database and then the application of Eq. 1 to highlight the network gatekeepers. Figure 2a shows in the inset the color scale, which was applied in Fig. 2a,b. The minimum degree (number of connections of a node) value is 1 which is yellow in color, the most connected nodes are in navy blue purple whose value is 75. Figure 2a shows in red the connections that every gatekeeper has, and they are marked by bigger yellow circles. It can be observed that all of them are connected to other nodes, but they, at first glance, do not appear to be of importance because of the few connection they have.

In Fig. 2b can be observed in red, not only the gatekeeper's connections but also the connections of the first connected nodes and how they have much more connections which comprises most of the network. The nodes at which gatekeepers are connected are hubs due to the highly connections they have.

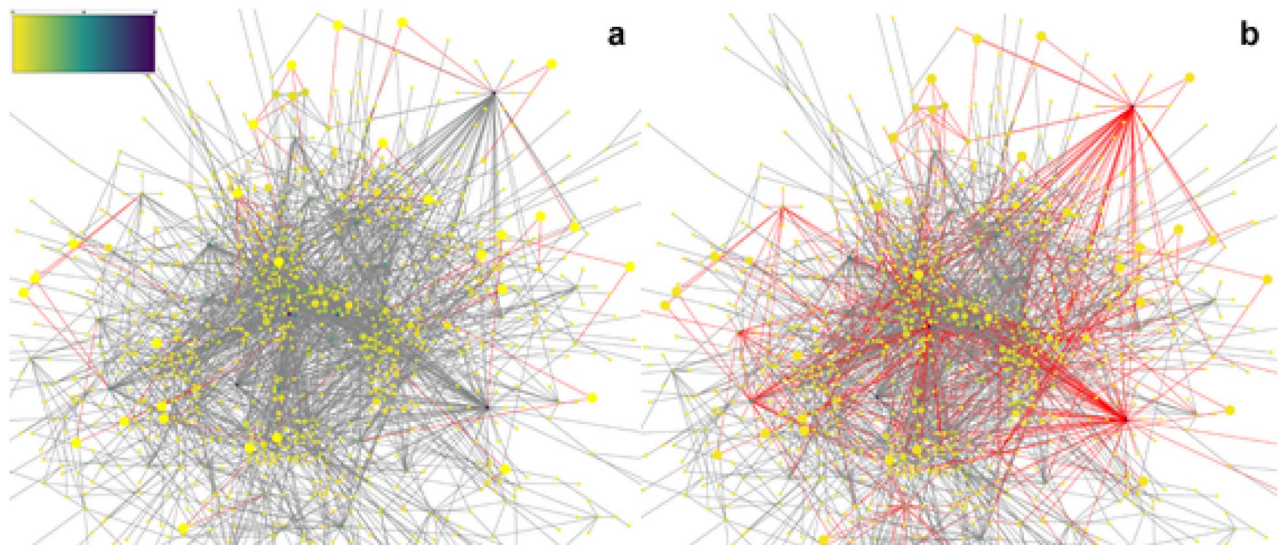


Figure 2. DGE-PPI network. **(a)** It represents a gathering and merging of the DGE analysis results and the Human Protein-Protein Interaction network from the Mentha Database. Red lines represent the connections of the gatekeepers. Color scale is presented in the left top corner. It represents the color scale applied to show graphically the values in the centrality measure for each node in the networks. **(b)** The red lines represent the connections of the gatekeepers plus the connections of its first neighbors (direct connected nodes to gatekeepers). The red lines comprise most of the network connections. **(a,b)** The proteins with less connections are marked in yellow and the most connected proteins are marked in navy blue purple. The size of each node represents the value of the clustering centrality measure; The bigger, the more value it has.

In Fig. 3, a zoom of the graph of Fig. 2 is presented and each node is tagged with its HGNC name tag. It can be observed to which nodes some of the gatekeepers are connected. For example, a connection (red) to MYC protein (purple) can be observed. This protein is considered as an oncogene frequently associated with poor outcomes; a gatekeeper is linked to other proteins as MEOX2 whose possible function in some cancers is to be a suppressor gene^{22,23}. Every one of the gatekeepers are linked to highly connected nodes which have relevant biological functions.

Table 1 summarizes in a list the results of the network analysis and the application of the machine learning algorithm in the GEO data set. A list of 26 gatekeepers' proteins is displayed in the first column which were the proteins with a clustering centrality of 1 obtained by the network analysis. This set of genes were used as an input list for the machine learning algorithm in which the results showed that they can be used to identify each carcinogenic stage with great accuracy. Second and third columns are two lists that contain a reduced set obtained only from the machine learning algorithm to classify each sample into its corresponding stages. Second column contains the probe tag used by the chip. Using two different methods we obtained two list of potentially gene sets that could be used as an aid to help classify samples and whose biological functions denote their potential use as targets for therapy. Further experiments are needed to probe its potential use.

An enrichment test was performed using the gatekeepers list to discover the main pathways associated with them as shown Table 2. The first characteristic is that every category is overrepresented, which means that in each presented category there are more genes from the input list than it can be expected (using as reference the *Homo sapiens* REFLIST) and most of the processes are involved in cell cycle-related specially in mitosis.

The next step was to search in distinct databases the list of genes obtained by the machine learning algorithm. We selected two pathway databases: the reactome pathways and the KEGG pathways. In Table 3, it can be observed a list of 8 genes for which information was available. The first column corresponds to its name, the second column to the related pathways in Reactome and the third to KEGG pathways. Some of the related pathways are usually altered in some types of cancer as Beta-catenin independent WNT-signaling, SMAD2/SMAD3, tight junction, ABC transporters, etc.²⁴⁻²⁷.

A network was made based in the results obtained from the machine learning algorithm first glance which comprised approximately 800 genes. It was observed a big component (when a significant proportion of the nodes in a graph are connected) created by some nodes as seen in Fig. 4a.

An interesting characteristic of the identified network by the machine learning algorithm was that some of the network gatekeepers identified by the DGE analysis were connected to this big component as seen in Fig. 4b. Some biological functions of some nodes in this network are well known to be relevant for cancer progression as PTEN which is a tumor suppressor altered in some types of cancer, as well as others like MCL1 which is an anti-apoptotic protein altered in some types of cancers. Also, MCL1 is being studied as a target for cancer patient treatment in small cell lung cancer^{28,29}. FAR1 is observed to play an essential role in the production of ether lipids/plasmalogens whose synthesis requires fatty alcohol. ABCA1 catalyzes the translocation of specific phospholipids from the cytoplasmic to the extracellular/luminal leaflet of membrane coupled to the hydrolysis of ATP. In cancer it was observed that its inhibition plays an important role for cancer survival due to an increase of mitochondrial

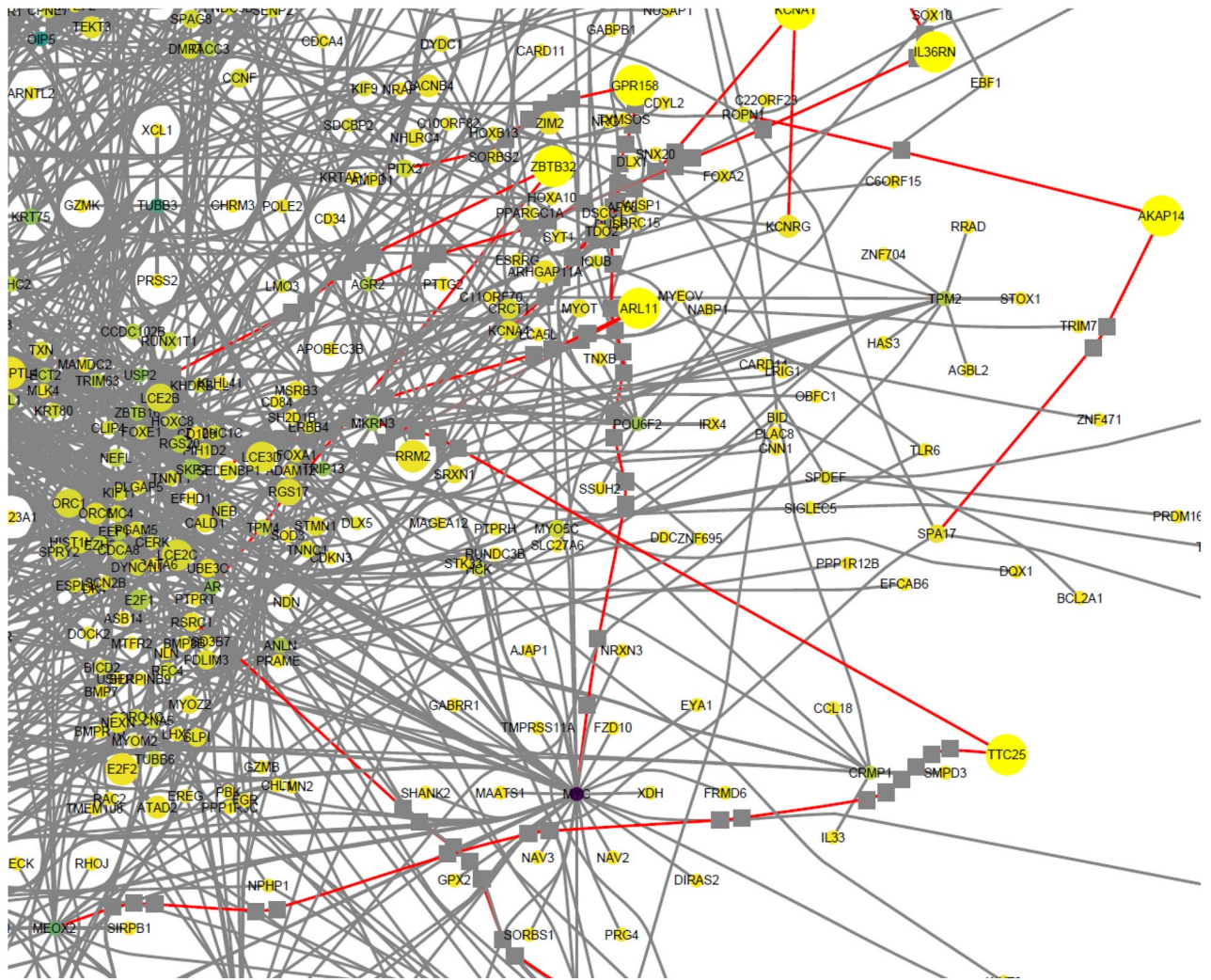


Figure 3. DGE-PPI network zoom. It can be observed with more detail some nodes with their respective connections. Red lines denote the connections of the network gatekeeper (nodes with a clustering centrality of 1) and some of the pointed nodes, in darker color, are associated to highly connected nodes.

cholesterol. The function of DMRT3 is not clear. It is thought to function as a transcription factor. In a study of lung cancer, the dysregulations of DMRT3 along with other two proteins were considered specific for lung squamous cell carcinoma³⁰. AAK1 is a kinase that participates in the regulation of clathrin-mediated endocytosis. It was discovered that in β -Catenin-dependent WNT signal a negative feedback loop is created by its expression. ASF1B is a histone chaperone which facilitates histone deposition, exchange, and removal during nucleosome assembly/disassembly; in cervical cancer it was observed that it functions as an oncogene accelerating cancer cells proliferation. APOC1 functions as an inhibitor of lipoprotein binding to the low-density lipoprotein (LDL) receptor. In gastric cancer it was proposed as a potential diagnostic and prognostic biomarker; in colorectal cancer evidence points out to have a promoting role in carcinogenesis. ADRA1B is an alpha-adrenergic receptor whose action is mediated by association with G proteins; in gastric cancer it was found a methylation promoter and it could be frequently involved in development and gastric cancer progression. These mentioned proteins are other examples of gene protein products whose functions are or could be involved with cancer disease^{30–37}.

Discussion

Lung cancer is the deadliest type of cancer, most of the diagnosed cases are made in the last stages of the disease and there are little available treatment options which could have an important effect. Small cell carcinoma of the lung comprehends a great part of all lung cancers. Our present results provide a better comprehension of the underlying components of the disease. The detection of genes and proteins that could be implicated in the carcinogenic process is urgently needed to provide better options for treatments and diagnosis. Herein, we performed a thorough search of genes and proteins that could be used to offer better treatments and diagnosis options. We made a comprehensive analysis of all the carcinogenic process and observed that some set of genes could be used as an aid for small cell carcinoma of the lung stage classification. We employed two pathways to identify relevant genes for diagnosis. The first was based in a classic method as DGE analysis with the aid of more

Gatekeepers (HGNC tags)	Probe tag Machine learning	HGNC tag for probe tag or genebank annotation
TTC25	A_23_P126803	ARPC5
SERPINA5	A_23_P216649	ABCA1
CENPL	A_23_P408865	Homo sapiens cDNA FLJ20700 fis, clone KAIA2250
ASF1B	A_23_P428366	HORMAD2
ZBTB32	A_23_P58009	C3orf52
GPR158	A_24_P100535	SYT15
RMI2	A_24_P141804	TMTC3
HSPB7	A_24_P239177	MUC4
ADRA1B	A_24_P515866	RBM6
GINS2	A_24_P542364	CALM1
APOC1	A_24_P59278	DSTYK
GINS1	A_24_P925678	PRG2
CENPK	A_24_P937366	**
KCNA1	A_32_P213091	LOC440338
PI3	A_32_P429083	LOC441621
ATP6V0D2		
ALS2CR12		
IL36RN		
KIF26B		
SPC25		
ARL11		
UBXN10		
LUM		
COTL1		
RYR3		
CENPI		

Table 1. Gene list from network gatekeepers and machine learning algorithm. Some Id are labeled as ** which means is a Missing Id. The first column corresponds to Gatekeepers list with 26 genes and the second column to the probe tag id in the microarray chip for 15 genes found with the machine learning method. The third column are the HGNC tags for each probe id of the second column. Second and third columns list finished when blank fields were present.

recent techniques as network analysis with a novel concept as the network gatekeepers which are encountered by using clustering centrality. DGE analysis was used as an exploratory analysis to look for possibly patterns in the gene expression for each stage. Although a general panorama of the carcinogenic process was obtained, we wanted to summarize it into a small meaningful set of genes with high involvement in cancer development. To make this possible we used the output data of the DGE as the input for the network analysis and then search for the network gatekeepers. The other pathway was based in another cutting-edge technique, machine learning algorithms. Hitherto, machine learning applications on cancer have been for assessing cancer prediction and prognosis³⁸. These results are based on the analysis of a wide set of variables including biomarkers and clinical factors such as age, location and type of cancer, and size of tumor^{39–41}. The results presented in here are not intended for cancer prevention or survivability directly, rather they provide a set of specific genetic biomarkers whose analysis can lead to an immediate diagnosis about the stage of development of small cell carcinoma in a patient. The analysis of the proposed genetic biomarkers can differentiate even the earliest stages of cancer development and lead a physician to administer the required treatment when the probabilities of survivability of the patient are higher. For each method we found a set of genes which we proposed to be useful as an aid for stage classification and due to the important biological roles in which they are involved they could also be useful for further validation as possible targets for treatment. The biological roles of the gatekeepers proposed set are marked as cell cycle regulation, DNA-repair breaks, nucleosome assembly and processes that occur in the mitotic phase. It was first observed in the gatekeeper network that most of them exhibit scarce connections but their first neighbor nodes which are directly connected are hubs (highly connected nodes). This along with the processes they are involved may permit an access for prior processes regulated by the hubs. It is known that network hubs are of high importance for network stability, but in this work, we are observing that it can be of great importance to use the network gatekeepers as a measure to find key components in a biological context. The machine learning algorithms are usually used in other fields to improve the understanding of a wide variety of processes. In the case of cancer its aim is to find new targets and possible key proteins that regulate cancer. We found a reduced set of genes that can be used for stage classification in a set of microarray data and this also can be done with the set of gatekeepers. The biological functions of each of the identified genes are relevant for normal stages and as previously observed for cancer development. For example, in the case of the reduced set

Reactome pathways	Homo sapiens—REFLIST (20,851)	Client Text Box Input (27)	Client Text Box Input (over/under)	Client Text Box Input (FDR)
Unwinding of DNA (R-HSA-176974)	12	2	+	2.07E-02
Deposition of new CENPA-containing nucleosomes at the centromere (R-HSA-606279)	54	3	+	1.23E-02
Nucleosome assembly (R-HSA-774815)	54	3	+	1.12E-02
Amplification of signal from unattached kinetochores via a MAD2 inhibitory signal (R-HSA-141444)	92	4	+	1.55E-02
Amplification of signal from the kinetochores (R-HSA-141424)	92	4	+	7.76E-03
Mitotic Spindle Checkpoint (R-HSA-69618)	108	4	+	7.15E-03
EML4 and NUDC in mitotic spindle formation (R-HSA-9648025)	114	4	+	7.03E-03
Chromosome Maintenance (R-HSA-73886)	90	3	+	3.11E-02
Resolution of Sister Chromatid Cohesion (R-HSA-2500257)	122	4	+	7.59E-03
RHO GTPases Activate Formins (R-HSA-5663220)	135	4	+	8.37E-03
Separation of Sister Chromatids (R-HSA-2467813)	185	4	+	1.85E-02
Mitotic Anaphase (R-HSA-68882)	193	4	+	2.00E-02
Mitotic Metaphase and Anaphase (R-HSA-2555396)	194	4	+	1.89E-02
Mitotic Prometaphase (R-HSA-68877)	198	4	+	1.91E-02
Cell Cycle Checkpoints (R-HSA-69620)	270	5	+	7.94E-03
Cell Cycle, Mitotic (R-HSA-69278)	495	6	+	9.07E-03
Cell Cycle (R-HSA-1640170)	600	7	+	6.91E-03

Table 2. Gatekeepers: Enrichment test-Reactome pathways. Main reactome pathways are shown if False discovery rate value was less than 0.05.

Name	Reactome	KEGG
ARPC5	EPH-Ephrin signaling	Tight junction, Regulation of actin cytoskeleton, Bacterial invasion of epithelial cells
ABCA1	Regulation of lipid metabolism by PPARalpha	ABC transporters
HORMAD2	Recruitment and ATM-mediated phosphorylation of repair and signaling proteins at DNA double strand breaks Processing of DNA double-strand break ends Nonhomologous End-Joining (NHEJ)	Null
C3orf52	SMAD2/SMAD3:SMAD4 heterotrimer regulates transcription Complex I biogenesis	Null
TMTC3	Reelin signalling pathway	Null
MUC4	O-linked glycosylation	Null
CALM1	Beta-catenin independent WNT signaling RAS processing RAF/MAP kinase cascade Signaling downstream of RAS mutant Signaling by RAF1 mutants	Null
PRG2	Neutrophil degranulation	Asthma

Table 3. Machine learning selected genes: Reactome and KEGG pathways involved. The pathways that could be related to squamous cell carcinoma of the lung are shown in Reactome column. Most of the genes do not have a pathway related to KEGG database, they are labeled as null. If there were more than three pathways available in either database just three pathways or less were selected when its biological function could be useful in cancer progression, growth, or maintenance. If just one pathway was available, it was written in the corresponding field. Null is used when no hits were found in the database. Only genes that do not appear in either database were not presented.

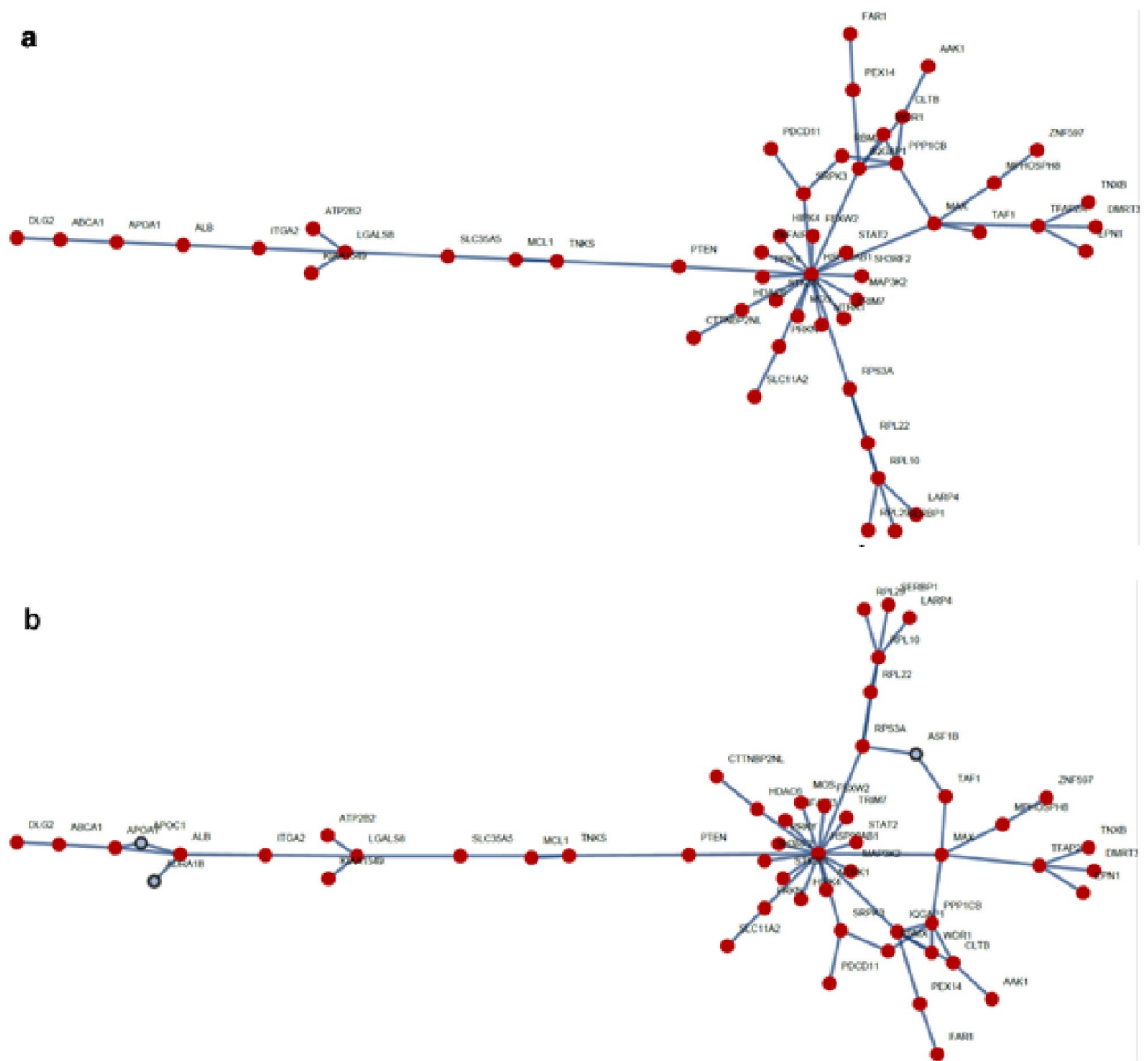


Figure 4. Big component network. (a) A network with 52 nodes is displayed. The network is a big component observed in the exploratory network analysis of a network created by the Machine learning algorithm. Nodes are displayed in red color; connections are in blue. (b) A network of 55 nodes of which 3 nodes are from the identified gatekeepers. Figures were made with the library NetWorkX of Python.

obtained with machine learning, *ARPC5* is a protein whose normal function is involved in EPH-Ephrin signaling and tight junction regulation and they are involved in cancer processes as adhesion, migrations, invasion or growth. This protein was recently proposed to be a prognostic biomarker for patients with multiple myeloma⁴². In the case of *ABCA1* is a protein whose inhibition promotes cancer progression³². Genes identified in the big component obtained from the machine learning first set was also analyzed and observed that some of them were previously studied in cancer and that their functions are involved with them. It is necessary to study these proteins in the context of squamous cell carcinoma of the lung, as it is known that the function is dependent of the type of tissue, microenvironment, and type of cancer. Our combined approach of DGE analysis plus the use of the metric of clustering centrality together with the application of machine learning algorithms, will facilitate the identification of relevant components in biological networks as the ones derived from cancer data.

Conclusions

We found a small set of genes possibly involved in the development of the disease. We propose two sets of genes which could help in the classification of tumor samples. These findings can increase the available options for experimental assays and molecular targets that could be used in novel treatment development. Although further experimental research is needed to validate their utility in the clinical setting.

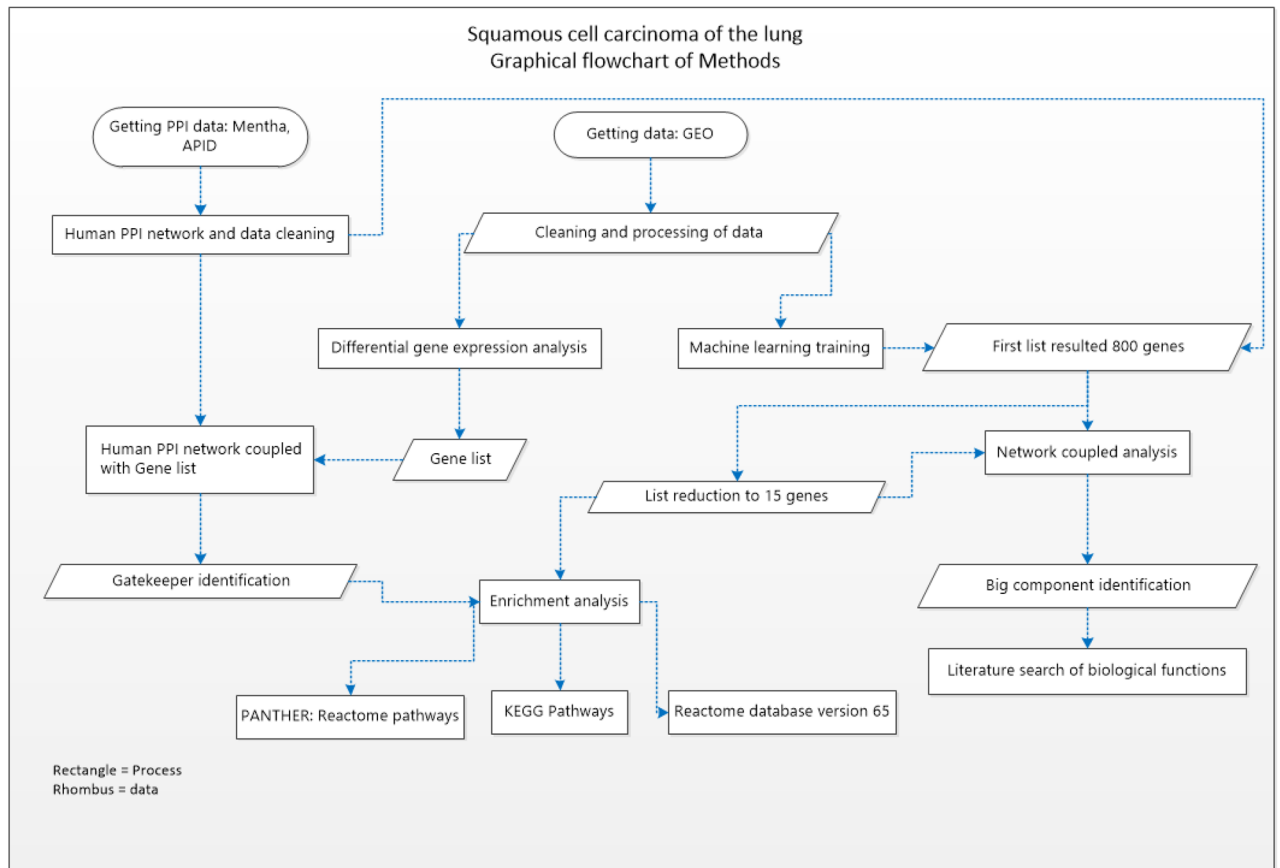


Figure 5. Workflow. A general panorama of the methodology and the databases. Figure made with Microsoft Visio.

Methods

A graphical flowchart that summarizes the methods and the data bases is shown in Fig. 5. Data collection was made by using various databases: Gene expression Omnibus for gene expression patients set, GEO accession: GSE33479, which comprises 122 patient samples representing the carcinogenic stages. Samples were divided as: 13 normal histology and normo-fluorescent, 14 with normal histology and hypo-fluorescent, those were grouped as the control group, 15 metaplasia samples, 13 mild dysplasia, 13 moderate dysplasia, 12 severe dysplasia, 13 carcinoma in situ, and 14 for squamous cell carcinoma of the lung. The gene expression platform was Agilent-014850 Whole Human Genome, Microarray 4 × 44 K G4112F. Processing and differential gene expression analysis were performed using R v3.5.1 software (<http://www.R-project.org>). Processed data retrieval was performed by GEOquery R package. Hgug4112a.db R package was used to annotate each gene ID to the data⁴³. Using limma package differential gene expression (DGE) analysis was used to compare each stage vs the normal. Limma package fits a generalized linear model before comparisons and then calculate a moderate t-statistic for each contrast^{44,45}. A p-value is obtained which is adjusted based in Benjamini and Hochberg False Discovery Rate correction^{44,46}. A list from the DGE was obtained for each comparison, results were merged to obtain a new list with all differentially genes. Full Human Interactome was downloaded from Mentha and APID database^{47,48}. Protein–protein interactions (PPI) level 0 data (all reported proteins pairs) was obtained from APID. Cleaning process for both networks was made using Cytoscape software (Networks for Figs. 2 and 3 were created using this software) which comprised: deletion of repeated interactions, deletion of protein interactions detected in other organism, deletion of self-loop interactions in proteins⁴⁹. Both databases are public and free to use. The merged list resulted from DGE (using a filter of $p < 0.05$ and $\text{Fold change} < -1.5$ & > 1.5) from the microarray data were coupled with Mentha PPI dataset which allowed to create a new network of PPIs which were used as a template to identify the network gatekeeper's proteins using clustering centrality measure Eq. (1). Mentha PPI data was used due to the better coverage of the genes that appeared in list of DGE analysis. To calculate clustering centrality measure we used the following Eq. (1):

$$C_i = \frac{2E_i}{k_i(k_i - 1)}$$

where C_i is the clustering coefficient of a node i and is defined as the fraction E_i of existing connections among its k_i nearest neighbors divided by the total number of possible connections.

Enrichment test. Statistical overrepresentation analysis was performed using PANTHER database for Reactome Pathways applied to Gatekeepers list using Fisher exact test. Raw p values were obtained. This value is the probability that the number of observed genes in each category occurred by chance. These p -values were corrected using False Discovery Rate by Benjamini-Hochberg. The reference list used was for *Homo sapiens*. Reactome database version 65 Released 2019-12-22 was used. In the case of machine learning gene list, it was not possible to use the PANTHER database due to the lack of information about them, none of them was annotated in the database, instead Reactome and KEGG Pathways database were used to perform individual searches of each gene in the list^{50–54}.

Data availability

The datasets generated analyzed during the current study are available in the GEO, Mentha and APID repository. GEO: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE33479> Mentha for human: <https://mentha.uniro.ma2.it/> APID for human: <http://cicblade.dep.usal.es:8080/APID/init.action>. The datasets generated during the current study are available from the corresponding author on reasonable request.

Received: 3 December 2020; Accepted: 3 February 2021

Published online: 01 March 2021

References

- Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **68**, 394–424 (2018).
- Gandara, D. R., Hammerman, P. S., Sos, M. L., Lara, P. N. & Hirsch, F. R. Squamous cell lung cancer: from tumor genomics to cancer therapeutics. *Clin. Cancer Res.* **21**, 2236–2243 (2015).
- Hashemi-Sadraei, N. & Hanna, N. Targeting FGFR in squamous cell carcinoma of the lung. *Target. Oncol.* **12**, 741–755 (2017).
- Singh, A. P., AdrianzenHerrera, D., Zhang, Y., Perez-Soler, R. & Cheng, H. Mouse models in squamous cell lung cancer: impact for drug discovery. *Expert Opin. Drug Discov.* **13**, 347–358 (2018).
- Drilon, A., Rektman, N., Ladanyi, M. & Paik, P. Squamous-cell carcinomas of the lung: emerging biology, controversies, and the promise of targeted therapy. *Lancet Oncol.* **13**, e418–e426 (2012).
- Heist, R. S., Sequist, L. V. & Engelman, J. A. Genetic changes in squamous cell lung cancer: a review. *J. Thorac. Oncol.* **7**, 924–933 (2012).
- Derman, B. A., Mileham, K. F., Bonomi, P. D., Batus, M. & Fidler, M. J. Treatment of advanced squamous cell carcinoma of the lung: a review. *Transl. Lung Cancer Res.* **4**, 524–532 (2015).
- Herbst, R. S., Morgensztern, D. & Boshoff, C. The biology and management of non-small cell lung cancer. *Nature* **553**, 446–454 (2018).
- Hirsch, F. R. *et al.* Lung cancer: current therapies and new targeted treatments. *Lancet* **389**, 299–311 (2017).
- Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).
- Albert, R., Jeong, H. & Barabási, A.-L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A.-L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
- Yildirim, M. A., Goh, K.-I., Cusick, M. E., Barabási, A.-L. & Vidal, M. Drug–target network. *Nat. Biotechnol.* **25**, 1119–1126 (2007).
- Breitkreutz, D., Hlatky, L., Rietman, E. & Tuszynski, J. A. Molecular signaling network complexity is correlated with cancer patient survivability. *Proc. Natl. Acad. Sci.* **109**, 9209–9212 (2012).
- Juarez-Flores, A. & José, M. Multivariate entropy characterizes the gene expression and protein-protein networks in four types of cancer. *Entropy* **20**, 154 (2018).
- Juarez-Flores, A. & José, M. Original Article Squamous cell carcinoma of the lung: gene expression and network analysis during carcinogenesis. *Int. J. Clin. Exp. Med.* **12**, 6671–6683 (2019).
- West, J., Bianconi, G., Severini, S. & Teschendorff, A. E. Differential network entropy reveals cancer system hallmarks. *Sci. Rep.* <https://doi.org/10.1038/srep00802> (2012).
- Deo, R. C. Machine learning in medicine. *Circulation* **132**, 1920–1930 (2015).
- Komura, D. & Ishikawa, S. Machine learning approaches for pathologic diagnosis. *Virchows Arch.* **475**, 131–138 (2019).
- Handelman, G. S. *et al.* eDoctor: machine learning and the future of medicine. *J. Intern. Med.* **284**, 603–619 (2018).
- Dreiseitl, S. & Ohno-Machado, L. Logistic regression and artificial neural network classification models: a methodology review. *J. Biomed. Inform.* **35**, 352–359 (2002).
- Chen, H., Liu, H. & Qing, G. Targeting oncogenic Myc as a strategy for cancer treatment. *Signal Transduct. Target. Ther.* **3**, 5 (2018).
- Tian, L. *et al.* Over-expression of MEOX2 promotes apoptosis through inhibiting the PI3K/Akt pathway in laryngeal cancer cells. *Neoplasma* **65**, 745–752 (2018).
- Tian, F. *et al.* Reduction in Smad2/3 signaling enhances tumorigenesis but suppresses metastasis of breast cancer cell lines. *Cancer Res.* **63**, 8284–8292 (2003).
- Voloshanenko, O. *et al.* β -catenin-independent regulation of Wnt target genes by RoR2 and ATF2/ATF4 in colon cancer cells. *Sci. Rep.* **8**, 3178 (2018).
- Salvador, E., Burek, M. & Förster, C. Y. Tight junctions and the tumor microenvironment. *Curr. Pathobiol. Rep.* **4**, 135–145 (2016).
- Sun, Y.-L., Patel, A., Kumar, P. & Chen, Z.-S. Role of ABC transporters in cancer chemotherapy. *Chin. J. Cancer* **31**, 51–57 (2012).
- Chen, C.-Y., Chen, J., He, L. & Stiles, B. L. PTEN: tumor suppressor and metabolic regulator. *Front. Endocrinol.* **9**, 338 (2018).
- Yasuda, Y. *et al.* MCL1 inhibition is effective against a subset of small-cell lung cancer with high MCL1 and low BCL-XL expression. *Cell Death Dis.* **11**, 177 (2020).
- Zhang, S., Li, M., Ji, H. & Fang, Z. Landscape of transcriptional deregulation in lung cancer. *BMC Genomics* **19**, 435 (2018).
- UniProt. <http://www.uniprot.org/> (2017).
- Smith, B. & Land, H. Anticancer activity of the cholesterol exporter ABCA1 gene. *Cell Rep.* **2**, 580–590 (2012).
- Agajanian, M. J. *et al.* WNT activates the AAK1 kinase to promote clathrin-mediated endocytosis of LRP6 and establish a negative feedback loop. *Cell Rep.* **26**, 79–93.e8 (2019).
- Liu, X. *et al.* ASF1B promotes cervical cancer progression through stabilization of CDK9. *Cell Death Dis.* **11**, 705 (2020).
- Yi, J. *et al.* Apolipoprotein C1 (APOC1) as a novel diagnostic and prognostic biomarker for gastric cancer. *Ann. Transl. Med.* **7**, 380–380 (2019).
- Ren, H. *et al.* Apolipoprotein C1 (APOC1) promotes tumor progression via MAPK signaling pathways in colorectal cancer. *Cancer Manag. Res.* **11**, 4917–4930 (2019).
- Noda, H., Miyaji, Y., Nakanishi, A., Konishi, F. & Miki, Y. Frequent reduced expression of alpha-1B-adrenergic receptor caused by aberrant promoter methylation in gastric cancers. *Br. J. Cancer* **96**, 383–390 (2007).

38. Cruz, J. A. & Wishart, D. S. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform.* **2**, 117693510600200 (2006).
39. Fielding, L. P., Fenoglio-Preiser, C. M. & Freedman, L. S. The future of prognostic factors in outcome prediction for patients with cancer. *Cancer* **70**, 2367–2377 (1992).
40. Cochran, A. J. Prediction of outcome for patients with cutaneous melanoma. *Pigment Cell Res.* **10**, 162–167 (1997).
41. Burke, H. B., Bostwick, D. G., Meiers, I. & Montironi, R. Prostate cancer outcome: epidemiology and biostatistics. *Anal. Quant. Cytol. Histol.* **27**, 211–217 (2005).
42. Xiong, T. & Luo, Z. The expression of actin-related protein 2/3 complex subunit 5 (ARPC5) expression in multiple myeloma and its prognostic significance. *Med. Sci. Monit.* **24**, 6340–6348 (2018).
43. Davis, S. & Meltzer, P. S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **23**, 1846–1847 (2007).
44. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).
45. Schurch, N. J. *et al.* How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?. *RNA* **22**, 839–851 (2016).
46. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
47. Calderone, A., Castagnoli, L. & Cesareni, G. mentha: a resource for browsing integrated protein-interaction networks. *Nat. Methods* **10**, 690–691 (2013).
48. Alonso-López, D. *et al.* APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks. *Nucl. Acids Res.* **44**, W529–W535 (2016).
49. Shannon, P. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
50. Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-scale gene function analysis with the PANTHER classification system. *Nat. Protoc.* **8**, 1551–1566 (2013).
51. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* gkz1031 (2019). <https://doi.org/10.1093/nar/gkz1031>.
52. Kanehisa, M. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
53. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**, D545–D551 (2021).
54. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **28**, 1947–1951 (2019).

Acknowledgements

Angel Juarez-Flores is a doctoral student from Programa de Posgrado en Ciencias Biológicas, Universidad Nacional Autónoma de México (UNAM) and a fellowship recipient from Consejo Nacional de Ciencia y Tecnología (CONACYT) (number: 775924) and this paper constitutes a partial fulfilment of the Graduate Program in Biological Science of the UNAM. Gabriel S. Zamudio is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and received a doctoral fellowship from CONACYT (number: 737920). MVJ was financially supported by DGAPA, PAPIIT-201019 UNAM, México. We thank Juan R. Bobadilla for his technical computer support.

Author contributions

Conceived the whole work: A.J.F., G.S.Z. and M.V.J. Gather and processing data A.J.F.; Performed calculations: A.J.F. and G.S.Z. Figures 1 and 4: G.S.Z.; Figs. 2 and 3: A.J.F.; Fig. 5: A.J.F. and M.V.J.; Conducted literature review: All authors; Wrote the paper: All authors. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.V.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021