Pontificia Universidad JAVERIANA Bogotá

UNIVERSITAS SCIENTIARUM
Revista de la Facultad de Ciencias
Vol. 10, 57-69

IDENTIFICACIÓN DE CRITERIOS PARA LA SELECCIÓN NATURAL DE RNA MENSAJEROS

J. González[1], F. Novoa[2], O. Acevedo[3], L. Lareo[1]

[1] Departamento de Nutrición y Bioquímica
[2] Departamento de Matemáticas
[3] Departamento de Física
. Facultad de Ciencias, Pontificia Universidad Javeriana, Cra. 7ª Nº 43-82. Bogotá, Colombia
l.lareo@javeriana.edu.co

RESUMEN

Para el estudio de la evolución biológica, se debió que establecer una correspondencia entre dos polímeros, que permitiera el entendimiento de cómo la información almacenada en los ácidos nucleicos daba lugar a proteínas específicas. Actualmente se sabe que las tripletas del código genético de mRNA establecen la correspondencia entre los polímeros y que la función de traducción es realizada por el tRNA asociado al ribosoma; sin embargo, el mecanismo por el cual una "frase" del "lenguaje" del código de los nucleótidos es escogido entre todos los arreglos posibles, para ser traducido es aún desconocido. El presente estudio asume que debido a la naturaleza de transferencia de información del proceso de codificación es posible que uno de los parámetros con los que se han caracterizado dichos fenómenos sea relevante para la selección. De esta forma se generará un modelo teórico que permita seleccionar entre una serie de factores de la teoría de información calculados *in silico* y estimación de algunos parámetros fisicoquímicos para el conjunto de todas las secuencias de nucleotídicas de mRNA probabilísticamente posibles, es decir, todas aquellas derivadas de las combinaciones entre los codones de los aminoácidos presentes en la secuencia de una proteína en particular, permitirá identificar el o los que regulan que sólo una de esas secuencias de mRNA sea la traducida en la naturaleza.

Palabras clave: modelos, mRNA, nucleótidos, simulación, selección natural.

ABSTRACT

For the study biological evolution, it is necessary to determine a correspondence between two polymers that make it possible to understand how the information stored in a nucleotide string codes for a specific protein. At the moment it is known that the triplets of the genetic code establish the correspondence among the polymers and that the translation function is carried out by the tRNA associated with the ribosome. However, the mechanism by which a sentence of the language of the code of the nucleotides code is chosen from among all the possible arrangements, for translation still unknown. The present study assumes that due to the nature of information transfer via the coding process, it is possible that one of the parameters among these which have characterized those phenomena would be relevant to the selection process. In this way a theoretical model will be generated that allows for selection among a series of factors

from information theory calculated *in silico* and the prediction from some physico-chemical parameters for the group of all the possible mRNA sequences, that is to say, all those derived of the combinations among the codons will allow the identification of the or those that regulate the selection of a unique sequence for those mRNA's that are translated in nature.

Key words: models, mRNA, nucleotides, natural selection, simulation.

INTRODUCCIÓN

La transferencia de información surgió como un principio universal, ayudando a determinar por medio de códigos especiales, modelos de pensamiento humano. Ésta, se convirtió en un concepto científico cuando se iniciaba la era de la comunicación electrónica. Un mensaje sólo transmite información cuando existe algún grado de incertidumbre, en el receptor, acerca de lo que el mensaje contendrá. Cuanto mayor sea la incertidumbre, mayor será el contenido de información transmitida. Estos conceptos genéricos se pueden aplicar a la transferencia de información genética. Para éste, el material consiste en un número reducido de símbolo simbolizados por cuatro letras: A (adenina), G (guanina), C (citosina) y T (timina) o U (uracilo). Este sistema de representación se asemeja al código binario de Shannon (Shannon, 1948) que sólo consiste en los dos dígitos 0 ó 1. Las teorías de información han sido desarrolladas para este tipo de lenguajes binarios y con esto se facilita aplicar los principios de esas teorías y establecer modelos para predecir y conocer cuánta información está contenida en una molécula de ADN.

La teoría de información de Shannon se puede aplicar a cualquier tipo de sistema informativo en que se envíen mensajes de una fuente a un receptor. Las secuencias de nucleótidos que constituyen el mRNA son consideradas en este estudio como la fuente del mensaje, y las cadenas de aminoácidos que conforman las proteínas y se encuentran en el extremo del "canal de comunicaciones" como receptor. En esta teoría, una buena comunicación en el mensaje debe estar codificado antes del envío y debe incluir cierto nivel de redundancia en el mensaje. Shannon demostró, en su segundo teorema para los ruidos de las señales, la existencia de códigos que mantienen un orden dentro del desorden general.

Actualmente es claro que las tripletas del código genético de mRNA establecen una correspondencia entre los ácidos nucleicos y las proteínas; sin embargo, aún es desconocido el mecanismo por el cual una "frase" del "lenguaje" del código de los nucleótidos es escogido para ser traducido entre todos los arreglos posibles. En el presente estudio se asume que dada la naturaleza de la transferencia de información es posible que uno de los parámetros con los que se han caracterizado dichos fenómenos sea relevante para la selección. Con base en este concepto se propone un modelo teórico para seleccionar factores de la teoría de información, como la entropía, el contenido de información según Shannon y según Chaitin-Kolmogorov, factores bioquímicos como los contenidos de mononucleótidos y dinucleótidos, las energías de formación y la estimación *in silico* de algunos parámetros fisicoquímicos. Estos procesos se realizaron para todas las secuencias de mRNA probabilísticamente posibles para cada una de las proteínas que se emplearon en la generación del modelo.

Antecedentes

De acuerdo con Lagerkvist (Lagerkvist 1980), no existe complementariedad química conocida entre el triplete de un codón

y el aminoácido correspondiente. Entre los primeros autores sobre el origen del código genético, Crick (1968) sugirió que éste surgió gradualmente, en escalones, de acuerdo con la utilidad del producto de la traducción o proteínas primitivas. Crick, adicionalmente, aplicó el principio Darwiniano, de selección natural, a las interacciones ácido nucleico - péptido. De acuerdo con este punto de vista, la acumulación de un péptido específico habría llevado a la acumulación de más oligonucleótidos específicos.

Una posible asignación al azar podría haber dado lugar a aminoácidos codificados por múltiples tripletes no relacionados; el codón XYN; donde X y Y son bases fijas y N cualquiera de las cuatro bases; con frecuencia codifica un solo aminoácido, es decir, los codones para cada uno de estos aminoácidos tienen la misma base en cada una de las dos primeras posiciones (glicina, alanina, valina, prolina y treonina). Cuatro de los seis codones para serina y leucina también presentan esta característica. Cuando N es restringido a U ó C, el codón XYN codifica un sólo aminoácido en cada uno de los casos. Así, parece que las dos bases iniciales del código pudieron haberse decidido en primer lugar en la evolución (Brener *et al*., 1976).

Otro aspecto que puede explicar la no aleatoriedad de la selección tiene que ver con la similaridad química entre aminoácidos relacionada con codones similares, lo cual, puede indicar que el reconocimiento de clases de aminoácidos pudo haber precedido al reconocimiento de un aminoácido individual. Una explicación a la transformación de las interacciones primitivas oligonucleótido-oligopéptido en un sistema de traducción funcional está en la actualidad sujeto a discusión.

Algunos autores han sugerido que los aminoácidos se asociaron al principio con sus codones o con sus anticodones, a través de un ajuste estereoquímico o compartiendo otras propiedades de complementariedad como hidrofobicidad o hidrofilicidad (Jungck 1987). Strickberger (1993) ha postulado una hipótesis alternativa, que sugiere que la universalidad del código genético es consecuencia de la supervivencia de solo uno de todos los posibles códigos ensayados en el pasado. Esta hipótesis implica que las relaciones primitivas entre aminoácidos y codones surgieron fundamentalmente al azar, y no por un apareamiento estereoquímico estricto. Esto implica que se deben haber producido un número elevado de códigos genéticos iniciales, cada uno de ellos utilizado por diferentes grupos. Sin embargo, con el paso del tiempo permaneció sólo uno y los otros acabaron extinguiéndose (Wong, 1975) así el código acabó siendo estable hasta entonces. Según Wong (1975), los individuos portadores del código que acabaron imponiéndose debían tener una o varias ventajas exclusivas, que les habrían proporcionado una superioridad competitiva importante. Una de estas ventajas podría haber sido el acoplamiento en fase entre la replicación del DNA y la división celular.

MÉTODOS

Para el presente estudio, se tomó una muestra de 10 proteínas seleccionadas considerando como principal criterio su longitud. La fuente de información para las secuencias, de las proteínas seleccionadas, fue el GenBank del NCBI (http://www.ncbi.-nlm.nih.gov). Esta información básica de las secuencias seleccionadas se presenta en la tabla 1.

Tabla 1
Identificación de proteínas de estudio

Organismo	ID proteína	ID nucleótido	Longitud (aa)
Rana lessonae	20146864	20146863	19
Gorrilla gorilla	20146866	20146865	19
Branta leucopsis	20146890	20246889	19
Homo sapiens	3002540	3002539	12
Ovis aries	1364225	1354	27
Sus scrofa	1377863	1377862	10
Cervus nippon	4432903	303530	13
Rattus norvegicus	5805063	5805062	15
Amytornis striatus	45478443	45478442	17
Neisseria gonorrhoeae	2662533	2662532	18

A cada una de estas secuencias proteicas se les realizó una traducción conceptual inversa, mediante el programa de Backtranslation del servidor de análisis proteómico Expasy (http://www.expasy.ch) La generación de todas las secuencias de mRNA probabilísticamente posibles se realizó empleando el software CoCoA System (Computation in Commutative Algebra) con base en el código genético universal.

Posteriormente se hicieron arreglos de las secuencias obtenidas para representarlas con la tercera base en forma genérica (N) y también se hicieron transformaciones a purinas y pirimidinas. Estos dos procesos, que se realizan manualmente, tienen el propósito de reducir el número de secuencias posibles para facilitar los procesos de análisis preliminares.

Propiedades fisicoquímicas

A todas las secuencias, antes de reducir el tamaño de muestra como se explicó anteriormente, se le realizaron las estimaciones de las siguientes propiedades fisicoquímicas: peso molecular, punto isoeléctrico, volumen, densidad, proporción de área hidrofílica, propensión para la donación y aceptación de puentes de hidrógeno, momento dipolar, capacidad calorífica, solubilidad y coeficiente de partición (logP). Estos cálculos fueron realizados empleando el Software Molecular Modeling Pro. (ChemSW, 2003).

Se realizó un estudio detallado del parámetro hidrofobicidad debido, básicamente, a las diferencias de valores encontradas previamente entre grupos de codones, lo cual, permitió generar algunas inferencias en cuanto a la asociación de los mismos con la posible selección del codón. Los datos comparativos se tomaron de Black & Mould (Black y Mould, 1991).

Posterior a la medición de los parámetros fisicoquímicos se determinó la frecuencia relativa de uso de codones de acuerdo con la base de datos Codon Usage de GenBank Release 141.0 (May 11 2004).

Otro parámetro considerado correspondió al conteo de mononucleótidos y dinucleótidos de cada secuencia así como el uso de codones en ellas.

Todos los análisis estadísticos se realizaron con el programa estadístico R (REF.).

El conteo permitió la identificación de diferencias en cuanto a la probabilidad de aparición simultánea de mononucleótidos en cada secuencia, se tuvo como punto de

referencia la probabilidad de aparición independiente de cada base en la secuencia real. En este punto cada frecuencia individual fue tomada para una secuencia simultáneamente y posteriormente sometida a la comparación con la secuencia real.

Obtención del conjunto de secuencias de mRNA de trabajo

Para las diferencias en cuanto a la probabilidad de aparición simultánea de mononucleótidos de cada secuencia, se tuvo como punto de referencia la probabilidad de aparición independiente de cada base en la secuencia real. En este punto, cada frecuencia individual fue tomada para una sola secuencia hipotética y posteriormente sometida a la comparación con la secuencia real.

Las coincidencias en cuanto a esa misma probabilidad para el grupo de secuencias posibles, permitió la identificación de secuencias equiprobables, simultáneamente, para cada nivel. Las secuencias homólogas en cuanto a la probabilidad con respecto a la real fueron seleccionadas y caracterizadas.

Estas secuencias equiprobables identificadas se sometieron a alineamientos múltiples con respecto a la secuencia real mediante el programa T-coffee (http://www.ch.embnet.org/software/TCoffee.html).

Finalmente, los datos fueron sometidos a un análisis de conglomerados mediante la técnica de redes neuronales del programa IBM DB2 Intelligent Miner for Data.

RESULTADOS Y DISCUSIÓN

De todos los parámetros intrínsecos de una molécula, la energía es la que en forma intuitiva, es considerada *a priori* como un factor de selección por parte de la naturaleza. En este trabajo se analizaron cálculos de la energía química para formar los enlaces, la energía potencial total y la energía de enlace para identificar si éstos han sido un factor dominante en la selección de los codones que existen actualmente en la naturaleza. Como variable dependiente de estos factores se tendrá el uso de codones en diferentes organismos como lo reporta la base de datos Kasuza (http://www.-kazusa.org.jp/codon) es decir, el conteo de codones de toda la secuencia genómica de los organismos y no sólo las regiones codificantes.

Los datos obtenidos corresponden a valores significativamente cercanos entre Purinas (A, G) y Pirimidinas (C, U) como se observa en la tabla 2, en donde, las más altas diferencias fueron consistentemente evidenciadas en codones con baja relación estructural.

Tabla 2
Energía de enlace

Base nitrogenada	Costo energético Kj/mol
Adenina	34.144
Guanina	34.103
Citocina	33.817
Uracilo	33.103

La anterior suposición es coherente con la relación de estos codones (codones sinónimos) que contienen las mismas bases en las posiciones primera y segunda para aminoácidos particulares y por consiguiente tienen valores cercanos para los parámetros energéticos calculados.

Los datos de las variables fisicoquímicas, que son consistentemente similares para grupos de codones, permiten suponer la elevada similitud entre las secuencias de mRNA codificantes teóricamente para la misma proteína, limitando de esta manera la elección de dichas variables como factor regulador o controlador de la elección de una o un grupo de secuencias codificantes para una misma proteína.

En el análisis de las posibles relaciones entre la energía total y el uso de codones en los organismos considerados no se encontró ninguna correlación significativa que indique que la naturaleza empleó este parámetro energético, al menos para los que existen actualmente, como se puede ver en la figura 1. Tampoco se evidenció ninguna relación significativa entre los valores energéticos de los codones es estrecho (promedio = -102,05 Kcal/mol, desviación estándar = 3,26), aún si se considera la variabilidad por estar constituidos por purinas y pirimidinas. Por el contrario el uso de codones presenta un amplio rango aun dentro de los mismos codones en los organismos considerados y entre codones dentro del mismo organismo.

De acuerdo con los datos obtenidos para las propiedades calculadas, los grupos de codones (sinónimos) muestran marcadas tendencias de similitud de acuerdo con las propiedades determinadas, estas tendencias permiten evidenciar un claro patrón de distribución en todos los casos para el parámetro de energías.

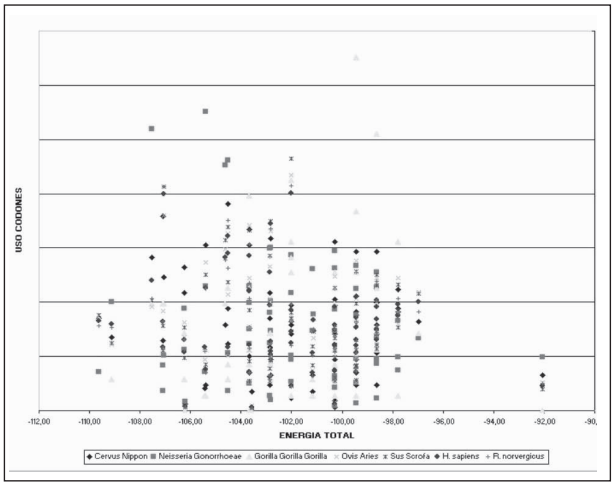

Figura 1.
Relación entre la energía de formación y el uso de codones en diferentes organismos de donde provienen las proteínas del estudio

Dada la selección separada de cada propiedad fisicoquímica, se puede inferir que el patrón de asignación de codones a aminoácidos puede estar estrechamente relacionado con la energía, dadas las tendencias intergrupos y la similitud intra grupos de tripletes sinónimos codificantes para un aminoácido en particular; sin embargo, dada esta similitud intragrupos no es posible determinar como factores controladores o condicionantes, de la selección entre los codones de un mismo grupo, a los factores energéticos.

De acuerdo con los patrones presentados de distribución de uso de codones se observa que los codones sinónimos no son usados en la misma proporción. Los resultados muestran una frecuente elección de codones con una preferencia general A, T ó G, C lo que sugiere que hay una selección determinada en codones que son traducidos más eficiente, rápida y certeramente.

Los conteos para cada una de las secuencias, demuestran que la secuencia real tiene un contenido más elevado de las bases GC con respecto a las demás. Para este punto se puede afirmar que las letras del tercer codón son casi invariantemente G ó C. Una situación igual se presenta para la primera

posición del codón, el cual, presenta reiterativamente las mismas bases, mientras que GC se conserva relativamente baja para la segunda posición.

Para revelar si el contenido de GC es una variable que afecta el uso de aminoácidos se acudió a las tablas estandarizadas por Nakamura (1997), las cuales, aparentemente no muestran una correlación significativa en cuanto al contenido genómico de GC, así, una acumulación de GC en el tercer codón parece no tener un gran efecto en el uso de aminoácidos dado el grado de redundancia del codón.

Se realizó el estudio de la distribución de frecuencias de mononucleótidos y dinucleótidos en cuanto al contenido de purinas y pirimidinas y de las duplas G+C y A+T; sin embargo, los datos no arrojaron resultados contundentes respecto a la definición de un patrón que permita la distinción de la secuencia real respecto a las demás posibilidades, para el caso de mononucleótidos y contenido de purinas (G+A) y pirimidinas (C+T).

Para el caso del contenido de A+T y G+C, para cada una de las secuencias, se evidencia una distinción clara de la secuencia real en cada uno de los casos, lo cual, puede obedecer a que las mutaciones de A ó T, G ó C ocurren más frecuentemente que en cualquier otra dirección. Entre genes codificantes de proteínas, la tercera base del codón es aquella cuyo contenido de G+C tiene la correlación más alta.

El estudio del conjunto de datos con respecto a los conteos generados, permitió establecer una clara distribución diferente de las probabilidades de cada secuencia, asignando varias posibilidades a un conjunto de mensajes posibles (grupo de mRNAs para cada secuencia), en donde, claramente se observa que la redundancia hace desiguales las probabilidades, en lugar de emparejarlas sobre toda la gama de posibilidades que permite discriminar y caracterizar los subgrupos que comparten las probabilidades simultáneamente para cada mononucleótido.

En la tabla 3, se muestra el número de secuencias que comparten valores de probabilidad para cada mononucleótido con respecto a la secuencia realmente traducida en cada uno de los casos estudiados.

Tabla 3
Secuencias de mRNA totales y equiprobables para cada proteína en el estudio

Secuencias	Teóricamente probables	Simultáneamente equiprobables
U1	384	11
U2	9216	79
U3	258	5
U4	16	2
U5	16	2
U6	96	6
U7	96	5
U9	1152	11

Las coincidencias en cuanto a esa misma probabilidad para el grupo de secuencias posibles permitió la identificación de secuencias equiprobables (igual probabilidad) para cada nivel simultáneamente.

Las secuencias equiprobables identificadas correspondieron a grupos significativamente reducidos. De acuerdo con la alineación de cada uno de ellos con respecto a la secuencia real, se encuentra al patrón de redundancia como uno de los factores determinantes de la selección del código de información.

Para los grupos estudiados, se observa claramente que las mutaciones en la primera base son muy ocasionalmente neutrales como el caso de UUA?CUA, las cuales, codifican para leucina. Las mutaciones en la tercera base son frecuentemente neutrales.

De acuerdo con los alineamientos no optimizados para los grupos de secuencias que comparten probabilidad de ocurrencia en cada caso estudiado, se observa que la primera posición del codón muestra la más baja correlación seguida de la segunda posición y tercera de acuerdo con el grado de redundancia del codón. Esto es consistente con la idea de que los cambios mayormente probables recaen donde éstas son neutrales así:

Tercera base ? Segunda base ? Primera base

Se evidencia adicionalmente, que las sustituciones en la tercera base tienden a ser transiciones preferiblemente que transversiones, esto es que purinas tienden a sustituirse por purinas y pirimidinas por pirimidinas y que las bases preferiblemente seleccionadas por los grupos son G-C. De lo anterior es posible inferir que los codones con contenido G-C en la tercera base deben tener una alta probabilidad de traducción para una diversa gama de proteínas y por ende deben tener mayor acción génica.

Estos resultados han mostrado que este efecto no es sólo una lejana posibilidad teórica y que puede ser demostrado (como en este estudio) mediante el análisis del contenido de información en secuencias de interés; si no que, de acuerdo con lo anterior, es posible inferir que las características relevantes de la redundancia corresponden al número de mutaciones puntuales que la definen. Esta idea actualmente es apoyada por el hecho de que una fuerte mutación combinada con la neutralidad dejará a algunos codones completamente o casi inutilizables (Epstein 1966, Woese y Dugre 1966, Osawa *et al*., 1992).

A partir de la identificación de subgrupos (conjuntos de secuencias equiprobables), fue posible realizar un estudio detallado al factor de hidrofobicidad, obteniéndose que para los valores referidos (Black y Mould 1991), se tienen dos subdivisiones binarias como sigue: Purinas R=(A, G) y Pirimidinas Y=(C, U). De acuerdo con la propiedad de hidrofobicidad, los codones pueden ser subdivididos de la siguiente manera:

1. {RRR, RRY, YRR, YRY} (conjunto con baja hidrofobicidad o conjunto hidrofílico) y

2. {RYR,RYY,YYR,YYY} (conjunto con valores altos de hidrofobicidad o conjunto hidrofóbico).

Aunque la diferencia relativamente es baja, es significante si se observa en asociación con las transiciones y transversiones para los subgrupos de secuencias estudiadas, así, los valores varían de acuerdo con los cambios observados en los codones para las secuencias equiprobables seleccionadas.

El estudio de los cambios de bases para las secuencias equiprobables, muestra que las sustituciones se realizaron mayoritariamente entre codones que tienen una composición igual en cuanto a contenido de purinas y pirimidinas, aunque, la composición de las bases para los codones relacionados sea diferente, adicionalmente muestra que estas sustituciones presentan una evidente tendencia hacia los cambios por codones que presentan valores menores del parámetro evaluado (hidrofobicidad).

La figura 2, muestra algunas sustituciones representativas, de acuerdo con los alineamientos generados para secuencias equiprobables dadas, adicionalmente se evidencia que los cambios se realizan siempre hacia los codones con valores menores de hidrofobicidad o en su defecto hacia un valor igual para todos los casos.

El fenómeno presentado de acuerdo con los resultados, es apoyado por el hecho del surgimiento de la vida en un ambiente acuá-

tico, en el cual, la minimización del costo energético pudo haber dado preferentemente una selección de sustancias que presentarán a su vez características que permitieran una óptima asociación con el ambiente de interacción. Aunque esto es muy probable, los estudios y las inferencias desarrolladas en este punto para el presente estudio requieren de verificación experimental.

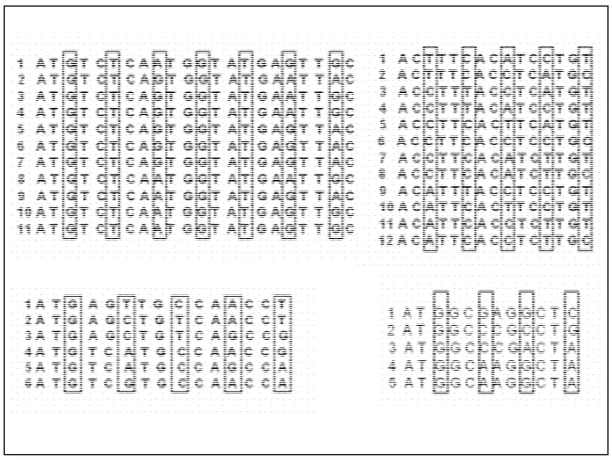

Figura 2.
Alineamientos de las secuencias de mRNA equiprobables.

Esta tendencia, fue consistente cuando se analizaron las medidas de polaridad presentadas por Ardell (1998), el cual, asume estas medidas (entre otros pará-metros) como unas de las fuerzas que pudieron haber estandarizado el código genético dadas diferencias en la fidelidad translacional por la posición del codón.

Comparación de los análisis con los valores de las secuencias de mRNA traducida.

El programa empleado (Intelligent Miner for Data) busca las características que se dan con más frecuencia y agrupa los registros relacionados de acuerdo con ello. El resultado de la función de agrupamiento muestra el número de conglomerados detectados y las características que los constituyen. El porcentaje poblacional al que corresponde cada conglomerado en cada uno de los casos se presenta en la tabla 4.

Tabla 4
Distribución poblacional de las secuencias en cada conglomerado generado por minería de datos

Secuencia	Conglomerado	Tamaño. Abs	Tamaño (%)
U1	3	23	5.99
U1	1	27	7.03
U2	6	987	10.71
U2	3	1193	12.94
U3	8	24	9.49
U3	5	48	18.97
U4	8	2	12.50
U4	6	4	25.00
U5	0	8	50.00
U5	5	4	25.00
U6	6	12	12.50
U6	7	6	6.25
U7	2	12	12.50
U7	8	10	10.42
U8	6	121	10.50
U8	3	134	11.63

Un análisis sistemático de la composición de los codones en diferentes grupos establece y corrobora el hecho de que la tercera base del codón marca una importante pauta para la selección del mismo en aminoácidos con grado de redundancia menor que seis. La distribución porcentual de las posiciones de las bases en los codones mutados refleja esta situación, dado el porcentaje mayoritario de cambio en la tercera posición para cada caso. Es necesario anotar que las posiciones primera y segunda para el codón que presentan significancia estadística, corresponden estrictamente a codones que presentan grado de redundancia igual a seis.

Este lenguaje usado que define la redundancia y es aplicado al código genético, permitió en este trabajo la identificación de algunos patrones que regulan la selección de una secuencia codificante entre todas las posibilidades equiprobables teóricamente y permiten por ende, generar un modelo de posible selección así:

1. El cambio en la primera base es muy ocasionalmente neutral.

2. Las mutaciones en la tercera base son frecuentemente neutrales, esto es, dicho cambio es predominante y no altera la codificación para un aminoácido particular entre codones sinónimos.

3. La primera posición del codón muestra la más baja correlación seguida de la segunda posición y tercera de acuerdo con el grado de redundancia del codón:

Tercera base >Segunda base >Primera base

4. Las sustituciones en la tercera base tienden a ser transiciones preferiblemente que transversiones.

5. Las sustituciones se realizan mayoritariamente entre codones que tienen una composición igual en cuanto a contenido de purinas y pirimidinas aunque la composición de las bases para la sustitución de los codones sea diferente.

6. Las bases preferiblemente seleccionadas para la tercera posición son G y C.

7. Invariantemente, el cambio en la tercera base se presenta si el grado de redundancia es menor o igual a 4.

$$\Delta B3 \text{ si } GR \leq 4$$

8. El cambio en la primera, segunda y tercera posición del codón se presenta simultáneamente en codones con grado de redundancia igual a 6.

$$\Delta B1; B2; B3 \text{ si } GR > 4$$

La distribución de las bases y sus mutaciones según su posición en el codón se presentan en la tabla 5.

Tabla 5
Distribución Porcentual de bases según su posición en el codón

Secuencia	Conglomerado	P1 %	P2 %	P3 %
U1	3	14.2	14.2	71.4
U1	0	14.2	14.2	71.4
U2	6	10.0	0.0	90.0
U2	3	10.0	0.0	90.0
U3	8	14.2	28.5	57.1
U3	5	14.2	28.5	57.1
U4	8	28.5	28.5	42.8
U4	6	28.5	28.5	42.8
U5	0	28.5	28.5	42.8
U5	5	28.5	28.5	42.8
U6	6	28.5	28.5	42.8
U6	7	28.5	28.5	42.8
U7	2	28.5	28.5	42.8
U7	8	14.2	28.5	57.1
U8	6	11.1	22.2	66.6
U8	3	11.1	11.1	77.7

CONCLUSIONES

Aunque de manera extensa son conocidos los detalles bioquímicos en cuanto a la implementación del código genético, la aplicación del mismo para la selección de una tripleta particular (de acuerdo con su grado de redundancia) que codifique y por ende haga parte de la secuencia de mRNA escogida por la naturaleza, entre todas las posibilidades probables que será traducida, es escasamente entendido. En particular, es muy poco conocido si la asignación de aminoácidos a tripletes es arbitraria o si son seleccionadas debido a procesos evolutivos.

Parte de esta ignorancia es debida a la imagen persistente del código congelado sugerido por Crick en 1968 (Crick, 1968).

Este estudio argumenta que el cambio en el patrón de redundancia, mientras se mantenga un conjunto de posibles aminoácidos constantes, es un factor determinante de la selección del código de información para la traducción proteica. Muestra adicional-mente, cómo las preguntas concernientes al impacto de los códigos para una única traducción también pueden ser cuestionadas en el contexto de la Teoría de Información mediante aplicaciones concretas de la misma teoría con asociaciones al código genético definidas entre secuencias tetranarias y posibles valores de variables involucradas en la definición de una posible solución. La expectativa correspondió al código genético, en el cual, patrones cuidadosamente escogidos como la redundancia pueden emplearse para generar un modelo.

Idealmente cada uno de los patrones de redundancia y el grado de neutralidad en la posición primera, segunda o tercera de cada base para la tripleta, podrían utilizarse para mejorar el modelo de selección propuesto. Esto, sin embargo, es un largo proceso debido al elevado número de posibilidades en secuencias teóricamente equiprobables de mRNA. En el presente estudio, se propone una disminución de datos mediante el cálculo de probabilidad de aparición independiente de mononucleótidos, definiendo así un óptimo local cuando el símbolo del código (en el sentido definido anteriormente) tiene una probabilidad mayor que todos los demás símbolos que pueden alcanzarse por mutaciones puntuales, así, fue posible contar el número de óptimos locales entre un conjunto de símbolos posibles. Tal posicionamiento constituyó el resultado de una función parcial obtenida con el análisis del grupo de símbolos posibles en posiciones no arbitrarias que corresponderán a las secuencias favorecidas con probabilidad simultánea correlacionada con la secuencia finalmente traducida.

Los resultados obtenidos con el estudio de los valores de hidrofobicidad para cada sustitución dada entre codones para las secuencias equiprobables, confirman los postulados anteriores dada una evidente tendencia hacia los cambios por codones que presentan valores menores del parámetro evaluado (hidrofobicidad). Este hecho adicionalmente es apoyado por el hecho del surgimiento de la vida en un ambiente acuático, en el cual, la minimización del costo energético pudo haber dado preferentemente una selección de sustancias que presentarán a su vez características que permitieran una óptima asociación con el ambiente de interacción. En la figura 3 se presenta una forma gráfica de visualizar los cambios de hidrofobicidad, siempre tratando de reducirla o al menos conservarla, con las mutaciones que se evidenciaron como criterio de selección con respecto a la tercera base del codón.

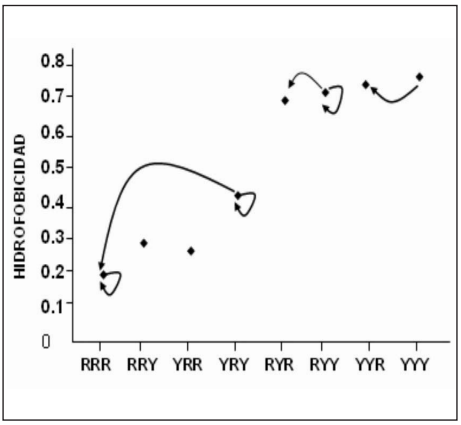

Figura 3.
Gráfica de la hidrofobicidad de los mRNA que han sido seleccionados por la naturaleza como codificantes en función de las mutaciones en su tercera base.
El valor de la hidrofobicidad se reduce y en los casos críticos se conserva en función del tipo de base en la tercera posición.

Estos resultados han mostrado que este efecto no es sólo una lejana posibilidad teórica y que puede ser demostrado (como en este estudio) mediante el análisis del contenido de información en secuencias de interés; así de acuerdo con lo anterior, es posible inferir que las características relevantes de la redundancia corresponden al número de mutaciones puntuales que la definen. Esta idea actualmente es apoyada por el hecho de que una fuerte mutación combinada con la neutralidad dejara a algunos codones completamente o casi inutilizables (Epstein, 1966; Woese y Dugre, 1966; Osawa *et al*., 1992).

Los análisis sistemáticos de la composición de la tercera base en el codón puede ser una herramienta muy útil para estudios genéticos relacionados con estudios de prospectiva funcional. Sin embargo, las bases mecanicistas de las ventajas en cuanto al contenido de GC en la tercera base del codón sigue siendo conjetural hasta el momento.

Dado que este estudio se desarrolló tomando como muestras proteínas de neuropéptidos, se puede establecer una clara relación en secuencias de longitudes cortas usando información derivada de patrones de redundancia que puede ser necesaria en la predicción de frecuencias relativas de nucleótidos para la selección de secuencias con una alta probabilidad de ser traducidas.

El modelo propuesto, genera evidencia tácita, en contra de la teoría de aleatoriedad en la selección de codones sinónimos para la selección de la secuencia de mRNA codificante. Considerando las secuencias de estudio, se encuentra que el cambio en cada una de las posiciones para las bases de la tripleta representa un direccionamiento crucial en la definición de la secuencia codificadora.

LITERATURA CITADA

ARDELL, D. 1998. On error minimization in a sequential origin of the standard genetic code. *J Mol Evol,* 47: 1-13.

BLACK, S. y MOULD D. 1991. Development of hydrophobicity parameters to analyze proteins which bear post or cotranslational modifications. *Anal Biochem,* 193: 207-209.

BRENER, A., CRICK, F. y KLUG, Y. 1976. A speculation on the origin of protein synthesis. *Origins of life,* 7: 389-397.

CRICK, F. 1968. The origin of the genetic code. *J Mol Evo*l, 38:367-379

CHEMSW. 2003. Molecular Modeling Pro. Version 5.2.4. Fairfield, CA 94534.

http//:www.expasy.org

http://www.kazusa.or.jp/codon/

http//www.ch.embnet.org/software/TCoffee.html.

JUNGCK, J. 1987. The genetic code as a periodic table. *J Mol Evol,* 11: 211-224.

EPSTEIN, C. 1966. Role of the amino acid "code" and of selection for conformation in the evolution of proteins. *Nature,* 210: 25-28.

LAGERKVIST, U. 1980. Codon misreading, a restriction operative in the evolution of the genetic code. *Am Sci*, 68: 192-198.

NAKAMURA, Y., GOJOBORI, K. y IKEMURA, T. 1997. Codon usage tabulated from the international dna sequence databases. *Nucleic Acid Res*, 25: 244-245.

OSAWA, S.; JUKES, T.; WATANABE, K. y MUTO, A. 1992. Recent evidence for evolution of the genetic code. *Microbiological Reviews*, 56: 229-264.

SHANNON, C. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27: 379-423, 623-656.

STICKBERGER, M. 1993. *Evolution*. Jones and Bartlett Publishers, Inc. Boston.

WOESE, C. y DUGRE, H. 1966. The molecular basis for the genetic code. *Proc Natl Acad Sci USA,* 55: 966-974.

WONG, J. 1975. A co-evolution theory of the genetic code. *Proc Natl Acad Sci*, *USA,* 72: 1909-1912.

Recibido: 12-05-2005
Aceptado: 12-09-2005