UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

INSTITUTO DE ECOLOGÍA

LA EMERGENCIA ESPONTÁNEA DE CÓDIGOS EN SISTEMAS PREBIÓTICOS : MODELAJE Y SIMULACIÓN

T E S I S

QUE PARA OBTENER EL GRADO ACADÉMICO DE

MAESTRO EN CIENCIAS BIOLÓGICAS (ORIENTACIÓN AMBIENTAL)

P R E S E N T A

JOSÉ AGUSTÍN MERCADO REYES

DIRECTOR DE TESIS: **DR. PABLO PADILLA LONGORIA**

COMITÉ TUTOR: **DRA. ALICIA NEGRÓN MENDOZA**
DR. ARTURO CARLOS II BECERRA BRACHO

MÉXICO, D.F. **DICIEMBRE, 2011**

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

INSTITUTO DE ECOLOGÍA

LA EMERGENCIA ESPONTÁNEA DE CÓDIGOS EN SISTEMAS PREBIÓTICOS : MODELAJE Y SIMULACIÓN

T E S I S

QUE PARA OBTENER EL GRADO ACADÉMICO DE

MAESTRO EN CIENCIAS BIOLÓGICAS (ORIENTACIÓN AMBIENTAL)

P R E S E N T A

JOSÉ AGUSTÍN MERCADO REYES

DIRECTOR DE TESIS: **DR. PABLO PADILLA LONGORIA**

COMITÉ TUTOR: **DRA. ALICIA NEGRÓN MENDOZA**
DR. ARTURO CARLOS II BECERRA BRACHO

MÉXICO, D.F. **DICIEMBRE, 2011**

UNAM
POSGRADO
Ciencias Biológicas

Dr. Isidro Ávila Martínez
Director General de Administración Escolar, UNAM
P r e s e n t e

Me permito informar a usted que en la reunión ordinaria del Comité Académico del Posgrado en Ciencias Biológicas, celebrada el día 27 de junio de 2011, se aprobó el siguiente jurado para el examen de grado de **MAESTRO EN CIENCIAS BIOLÓGICAS (BIOLOGÍA AMBIENTAL)** del alumno **MERCADO REYES JOSÉ AGUSTÍN** con número de cuenta **401052695** con la tesis titulada **"LA EMERGENCIA ESPONTÁNEA DE CÓDIGOS EN SISTEMAS PREBIÓTICOS: MODELAJE Y SIMULACIÓN."**, realizada bajo la dirección del **DR. PABLO PADILLA LONGORIA:**

Presidente:	DR. CARLOS GERSHENSON GARCÍA
Vocal:	DR. ALFONSO ARROYO SANTOS
Secretario:	DR. ARTURO CARLOS II BECERRA BRACHO
Suplente:	DR. LEÓN PATRICIO MARTÍNEZ CASTILLA
Suplente:	DRA. ALICIA NEGRÓN MENDOZA

Sin otro particular, me es grato enviarle un cordial saludo.

A T E N T A M E N T E
"POR MI RAZA HABLARA EL ESPIRITU"
Cd. Universitaria, D.F., a 30 de noviembre de 2011.

M. del Coro Arizmendi

DRA. MARÍA DEL CORO ARIZMENDI ARRIAGA
COORDINADORA DEL PROGRAMA

c.c.p. Expediente del interesado.

Edif. de Posgrado P. B. (Costado Sur de la Torre II de Humanidades) Ciudad Universitaria C.P. 04510 México, D.F.
Tel. 5623-0173 Fax: 5623-0172 http://pcbiol.posgrado.unam.mx

Agradecimientos

Quiero agradecer, en primer término, al Posgrado de Ciencias Biológicas, por permitirme realizar los estudios que culminan con esta tesis.

En segundo lugar, al Consejo Nacional de Ciencia y Tecnología, por el apoyo económico recibido a través de su beca para posgrado en la Convocatoria de Becas CONACyT Nacionales Agosto-Octubre 2009 (convocatoria número 290564, becario número 231056).

Finalmente, a mi tutor, el Dr. Pablo Padilla Longoria, y a los miembros del Comité Tutor, compuesto por el Dr. Arturo Becerra Bracho y la Dra. Alicia Negrón Mendoza.

Agradecimientos personales

A través del pequeño arco temporal que representa este posgrado he tenido apoyo invaluable de familia, amigos y colegas. La gratitud que por todos ellos siento no se puede expresar con palabras.

En primer lugar, y sobre todo, quiero agradecer a mis padres. El invaluable e irremplazable apoyo y cariño que me siguen dando, después de tantos años, hacen que la deuda que tengo con ustedes sea eterna. A ustedes dedico, una vez más, mi trabajo.

A Adriana, quien a pesar de no conocer este trabajo, ha hecho tolerable la inversión de tiempo y esfuerzo mediante rituales y costumbres (ttt).

A Álvaro, a quien he llegado a considerar parte de mi familia, con quien he compartido un espacio durante más de un año y que fue uno de los dos lectores tempranos de esta tesis; y a mis familiares cuadrúpedos (Drax y Mews) quienes fueron testigos presenciales de la escritura.

A mis amigos pasados y presentes, quienes me han acompañado en persona o en memoria, animado y tolerado, a lo largo de estos años. You know who you are.

A los docentes del posgrado, como Jorge Meave y Mark Olson, quienes a través de sus clases me enseñaron, simultáneamente, a ser maestro y alumno.

A Arturo Becerra y Alicia Negrón, quienes integraron mi comité tutor, y gracias a quienes esta tesis encontró su estructura y su camino.

A Alfonso Arroyo, Carlos Gershenson y León Martínez, quienes fueron (junto con los miembros de mi comité) los lectores de este trabajo. En ellos he notado esa envidiable e inusual costumbre de dejar caer una frase cargada de implicaciones, de la cual puede brotar un bosque de ideas.

Finalmente (*but in no way least*) al Dr. Pablo Padilla, que asesoró y dirigió esta tesis. Su guía tuvo ese extraño balance entre constricción positiva y libertad, ese punto medio en el que existe una labor creativa. Sin embargo, mi agradecimiento trasciende lo académico: también le agradezco las pláticas acerca de literatura o de música, las anécdotas compartidas y en general la amistad que me ha brindado a lo largo de casi cinco años de trabajo.

La nature est un temple où de vivants piliers
Laissent parfois sortir de confuses paroles;
L'homme y passe à travers des forêts de symboles
Qui l'observent avec des regards familiers

CHARLES BAUDELAIRE. *Correspondances*

Resumen

Después de discutir las implicaciones de la significación y de las relaciones simbólicas en los sistemas biológicos, presentamos un experimento informático para explorar la naturaleza del procesamiento de secuencias de DNA a través del código genético. Elegimos dicho sistema por su universalidad y su simplicidad relativa en comparación a otros tipos de semiosis biológica. Después de generar una población de códigos alternativos totalmente aleatorios, comparamos las diversas traducciones que realizan de las librerías de cDNA de cuatro organismos representativos de la biodiversidad conocida; llevamos a cabo la comparación a través de un algoritmo de compresión para tomar en cuenta no sólo el contenido infomativo *sensu* Shannon, sino lo posibles patrones internos. Los resultados hacen notar que las traducciones realizadas por el código genético estándar son significativamente más complejas que las de la gran mayoría de los otros códigos posibles, sin llegar a una incompresibilidad máxima. Finalmente, se analizan las implicaciones de dichos resultados; una de las más importantes es el efecto bidireccional que existe entre el mensaje y el código, lo cual parece apuntar a un origen común y sincrónico.

Abstract

After reviewing the implications of meaning and symbolic relationships in biological systems, we put forward an informatic experiment in order to explore the nature of the processing of dna chains by the genetic code. We chose said object of study because it is the most universal and simple semiotic system in the realm of biology. After generating a population of a thousand alternative, random codes, we compared the different translations of the genomes of four representative organisms. We analysed the obtained sequences through a compression algorithm to explore not only the informational content sensu shannon, but also to visualize their internal structure. Our results show that the standard genetic code translates sequences with an unusually high level of complexity, without being totally incompressible. This echoes several hypothesis in which a state near high complexity is needed for order to arise. Among the implications of these results, the most significative is the effect of the code on the genetic messages, and viceversa. This bidirectional effect can be evidential of a common, synchronic origin of information and translation.

Introducción

La biosemiótica es una rama novedosa y fértil de la biología. Desde su surgimiento, en las últimas décadas del Siglo XX, ha llamado la atención hacia la importancia del modo simbólico de funcionamiento de los seres vivos. Sin embargo, es tan heterogénea como su objeto de estudio: el rango de las investigaciones varía desde el comportamiento simbólico de los aspectos moleculares de la vida hasta el estudio semiótico de los ecosistemas. En lo referente a la complejidad de los sistemas simbólicos, los análisis varían desde un enfoque en la semiosis de códigos con equivalencias uno a uno, hasta los lenguajes animales, que dependen de una interpretación subjetiva. Por supuesto, en los puntos medios de este amplio espectro se encuentran la expresión genética y los mensajes intracelulares, la comunicación entre células distintas, el complicado control endócrino, la intepretación de señales físicas, los adaptadores utilizados para reconocimiento de factores del medio y un extenso *et cetera*.

Así, podemos afirmar que la biosemiótica constituye una rama un poco dispersa de la biología. Existen muchas otras, como la biología molecular, que dejan ver sus efectos en varios niveles del fenómeno de la vida; sin embargo, en ellas, tanto el objeto de estudio como las maneras de acercarse a él son precisos. En cambio, en la biosemiótica, no existe aún un canon de técnicas y delimitaciones que defina los objetos de estudio; en parte, esto se debe a la multiplicidad de disciplinas que interactúan en ella pero, sobre todo, se debe a la multiplicidad de voces y a la falta de tendencias definidas en el campo, las cuales únicamente en fechas muy recientes comenzaron a tomar forma. La presente investigación parte directamente de los intereses de la biosemiótica, y de la certeza de que el manejo simbólico de la información es una de las características que definen a la vida. Como argumentamos a lo largo del escrito, nos dedicamos a la semiosis que subyace en la totalidad del mundo biológico, es decir, el código genético. Al tomarlo como nuestro objeto de estudio, nos permite intentar definir ciertos conceptos básicos de la biosemiótica, con la simplicidad y la universalidad necesarias de un sistema modelo.

El trabajo se desarrolla en dos partes. En la primera, se expone un bosquejo de la historia de tal rama, así como de los conceptos fundamentales que se utilizan de ella, como sentido, significado, símbolo, e información. También se detalla las conexiones que existen entre el estudio de la semiosis y el de la información desde un punto de vista matemático. Además, se presenta como antecedentes las características conocidas del código genético, y cómo éstas pueden ser interpretadas como una guía para proponer un escenario acerca de su origen.

En la segunda parte se desarrolla el experimento informático que diseñamos para explorar algunas características del código genético, el cual únicamente representa un aspecto muy limitado de la semiosis viva. Dicho experimento es una propuesta novedosa para analizar la naturaleza de cualquier código: se traduce una cadena modelo de DNA, en este caso obtenida de las librerías de cDNA, utilizando códigos generados aleatoriamente. Mediante un algoritmo de compresión se obtiene un valor de complejidad de la secuencia, que es más completo que la entropía de Shannon y otros parámetros para medir información de mensajes. El resultado obtenido permite observar que la población de códigos aleatorios presenta una distribución que se mantiene constante sin importar el genoma utilizado; además, el código genético estándar produce consistentemente produce cadenas protéicas con complejidad elevada. Aun con las restricciones metodológicas que impusimos, creemos que los resultados obtenidos dan pie a una discusión extensa de varias facetas del sistema de estudio, y proponen implícitamente algunas preguntas para investigaciones futuras.

Ésta es una investigación acerca de la naturaleza actual del código genético, pero no podemos evitar incidir, una y otra vez, en el posible escenario de su origen y su evolución temprana. La vida es un palimpsesto, en el que las imágenes de eventos perdidos pueden ser vistas pálida pero incuestionablemente. La complejidad de la vida está formada por capa tras capa de accidentes históricos, de necesidades locales que han originado estructuras permanentes, de cambios bruscos o graduales en el interior o en el exterior de los organismos, de usos equívocos de sistemas que inicialmente tenían funciones distintas, de tendencias y prefiguraciones. Hablar acerca de alguna característica biológica es hablar, implícita o explícitamente, de su historia.

1. Símbolos de la vida

Breve historia de la biosemiótica

La variedad que existe en el *corpus* de la investigación biológica es comprensiblemente vasta. El fenómeno de la vida se expresa en diversos niveles espaciales, desde el molecular hasta el global. Estos niveles se pueden estudiar de manera independiente, o bien, correlacionarlos. Por ejemplo, es posible estudiar efectos moleculares en el desempeño ecológico de algún organismo, o la influencia de las condiciones de un ecosistema en el desarrollo ontogénico de un individuo: las combinaciones son virtualmente infinitas. Además la vida cuenta con un componente histórico cuyo estudio conforma el centro conceptual de la biología, como lo expresó Dobzhansky en su cita clásica: "Nothing in biology makes sense except in the light of evolution". (aparecida por primera vez en Dobzhansky, 1964, y la cual le dio nombre al famoso artículo de 1973). Esto permite relacionar filogenéticamente a los individuos mediante ancestros comunes, así como entender mejor el origen y el desarrollo de diversas características.

Sin embargo, existe un aspecto que se ha ignorado sistemáticamente en la mayoría de los estudios biológicos, al menos hasta hace unas cuantas décadas. Existe comunicación semiótica en los sistemas vivos en varios niveles. La más fácilmente percibida se encuentra en los animales: muchas formas de comunicación zoológicas trascienden el esquema de estímulo-respuesta, estableciendo relaciones simbólicas. Sin embargo, a medida que los estudios biológicos esclarecían los comportamientos de diversas señales, surgieron ejemplos de comunicación basada en moléculas que no se basaban en interacciones físicas. Con la elucidación del sistema molecular de almacenamiento genético (Watson y Crick, 1953) y el posterior descubrimiento del sistema ribosomal de traducción y el código genético, las relaciones semióticas fueron innegables. Una molécula (un aminoácido particular) era representada por otra (un triplete de nucleótidos), sin tener más enlace físico que la mediación de una molécula adaptadora (el RNA de transferencia). Aún más, dicha representación es completamente arbitraria: no hay ninguna necesidad física directa para que estén asociadas en el código genético.

Aún con la variedad de relaciones semióticas de ocurrencia natural en los sistemas biológicos, la síntesis explícita entre la ciencia de los signos y la ciencia de la vida tuvo que esperar a los trabajos de Thomas Sebeok; semiólogo de formación, exploró inicialmente distintas manifestaciones simbólicas de comunicación y comportamiento animal, las cuales agrupó bajo el nombre de "zoosemiótica" (Sebeok, 1965). Al surgir más información acerca de los procesos biológicos organísmicos, particularmente los del sistema inmune y endócrino, así como los procesos genéticos de comunicación molecular, señaló la necesidad de trabajar dentro de un campo al que denominó "endosemiótica" (Sebeok, 1985). Al incluir en su gran síntesis diversos aspectos de la vida que se manifiestan como relaciones semióticas, fundó finalmente el campo de la biosemiótica (Sebeok, 1986). Creo necesario mencionar dos aspectos importantes de la obra de Sebeok. Primero, fue estricto en la inclusión de fenómenos que clasificaran como simbólicos. Por ejemplo, excluía a los experimentos de comunicación gráfica o verbal de gorilas y otros primates como fuera del campo de la biosemiótica, pues los consideraba como una respuesta conductual a índices o iconos, no a símbolos (señalado en Favereau, 2007). Segundo, aunque tanto su obra publicada como su organización de conferencias, congresos y colaboraciones fueron de importancia central para la fundación de la biosemiótica, no fue el primero que notó el evidente fenómeno simbólico en el mundo biológico. Sin embargo, sí lo fue en consolidar a un grupo masivo de científicos para trabajar en el campo de estudio naciente.

Una de las grandes contribuciones teóricas de Sebeok fue su preferencia explícita por la semiótica de C.S. Peirce. Este tipo de acercamiento a los símbolos es distinto a la semiótica de Ferdinand de Saussure (Saussure, 2000), cuyo núcleo bipartita se define por la relación entre significado y significante. Según Saussure, la relación semiótica sólo se establece entre una palabra (el significante) y el concepto al cual señala (el significado). Alternativamente, es permisible la misma relación entre palabras de distintos lenguajes, conectadas por un significado común. Sin embargo, como lo señala Umberto Eco (Eco, 1968), esta definición del fenómeno semiótico es muy pobre, porque excluye de su campo de acción posibles signos como los que aparecen en la teoría de la información. La semiótica debe poder trabajar con símbolos y signos abstractos, e incluir fenómenos tales como la notación en la partitura musical, en cuyo caso sería difícil hablar explícitamente de un "significado conceptual", como lo requiere la teoría de Saussure. En este punto reside el

empobrecimiento de esta teoría: Saussure siempre hace equivalente el significado con una imagen mental, por lo que la semiótica se restringe a los procesos cognitivos humanos, e incluso a un conjunto limitado de ellos.

C.S. Peirce propone una definición mucho más amplia y completa. Además de los elementos de Saussure (que Peirce llama "un signo y su objeto"), postula la necesidad de un tercer elemento, que puede ser llamado "interpretante", que no representa a un intérprete, sino a una referencia a un código establecido (Eco, 1968; Chandler, 2007). Éste constituye la mediación entre el signo y su objeto, y hace que el primero sea representación del segundo para el destinatario. Los tres elementos son *suficientes y necesarios* para definir la relación simbólica, y aunque tienen la misma importancia para el sistema, es la acción del intepretante lo que establece el inicio del fenómeno semiótico. En palabras de Peirce, "*nothing is a sign unless it is interpreted as a sign"* (citado en Chandler, 2007). Esta definición del signo puede parecer tan antropocéntrica como la de Saussure, y resultar igualmente inútil para el estudio de los fenómenos biológicos. Sin embargo, Peirce tiene una concepción del universo que se acomodó perfectamente a los fines de Sebeok. El universo es pansemiótico y cualquier cosa tiene la potencia de ser un símbolo. Según él mismo, "the universe is perfused with signs, if it is not composed exclusively of signs" (citado en Nöth, 1995). El acercamiento a la biosemiótica, basado en las propuestas de Peirce y Sebeok, es conocido como la escuela Copenhagen-Tartu, lugares de residencia de dos de los centros de biosemiótica más importantes de Europa.

Existe una propuesta más, cuyo exponente original, y el más importante hasta ahora, es Marcello Barbieri. Aunque se relaciona cercanamente con la visión de Sebeok, tiene diferencias sustanciales. De nuevo se favorecen las ideas centrales del sistema semiótico tripartita de Peirce, pero en vez de relacionar significado con significante a través de un intérprete, se les relaciona mediante un *código natural (*Barbieri, 1982, 2003). Esta diferencia, aparentemente sencilla, acarrea una serie de profundas consecuencias en la visión de la semiótica y, en último término, en la manera de investigar la vida. En la escuela de Copenhagen-Tartu, de manera explícita o implícita, se equipara la arbitrariedad inherente a la semiótica con interpretabilidad. En cuanto existe un sistema semiótico, emerge la hermenéutica, y el intérprete es una condición necesaria (y en algunos casos, suficiente) para el surgimiento del comportamiento simbólico. Esto se debe a que tanto

Sebeok como sus seguidores (por ejemplo, Emmeche y Hoffmeyer, 1991; Hoffmeyer, 1998; Kull 1999) se ocupan de comportamientos en los que algún agente recibe cierta información, la procesa, infiere algo acerca de su medio y lleva a cabo cierto comportamiento. Este tipo de funcionamiento se encuentra en los sistemas más complejos del mundo biológico, como el procesamiento de estímulos en organismos completos, la transformación de señales eléctricas en acciones a través del sistema nervioso e incluso la emergencia de símbolos en ecosistemas. En ellos, la semiótica en efecto está indiscutiblemente unida a interpretación. Por su parte, Barbieri voltea su mirada hacia sistemas simbólicos en los que la significación de una cosa por otra distinta no está sujeta a la hermenéutica, y en donde las relaciones simbólicas existen pero son fijas, tales como los mensajes transmitidos por receptores celulares, la transformación de genotipo en fenotipo y, sobre todo, el código genético.[1]

La posición de Barbieri es, pues, una oposición al holismo absoluto. Barbieri trata de salir del nivel organísmico, señalando sistemas de reglas que funcionan como las máquinas, o al menos con la lógica mecanística propia éstas, logrando que el análisis de los diversos componentes del sistema sea posible. Así, parece conseguir cosas opuestas: no es reduccionista, pero aboga por la importancia de los elementos que componen al sistema. No es fisicalista, pues considera que el código genético no es sólo una metáfora, pero provee de métodos para acercarse a los sistemas semióticos de manera física. Por esto, la posición de Barbieri y el campo de estudios semióticos que sus investigaciones han abierto son el modelo a seguir del presente trabajo; nuestra ambición, que tal vez no logremos del todo, es encontrar un balance entre vistas opuestas que, sin embargo, conviven en un mismo ente o proceso del mundo físico. Para intentar acercarnos a este objetivo, es necesario definir algunas de estas oposiciones o problemas que se encuentran a cada paso en el estudio de la semiótica y de los símbolos naturales.

1 Para una extensa discusión acerca de la historia de la biosemiótica desde Grecia antigua hasta los avances de Marcello Barbieri, es conveniente consultar Favereau, 2007

El problema materia-símbolo

Los sistemas simbólicos conllevan un problema ineludible, que se hace especialmente evidente en los símbolos que tratamos de estudiar en el presente trabajo, es decir, los "códigos naturales" o "códigos orgánicos" (Barbieri, 2003). Todo símbolo tiene como base un sustento material. Por ejemplo, los símbolos que se crean en el proceso cognitivo cuya base física radicaen el cerebro, y, más específicamente, los impulsos eléctricos transmitidos entre neuronas. El código genético funciona como un sistema de relaciones semióticas, pero éstas se representan físicamente en la colección de moléculas que intervienen en la elaboración de proteínas. Esta doble identidad de los sistemas simbólicos (como relaciones abstractas entre elementos y como objetos físicos) resulta en dos modos de existencia, diferenciados pero complementarios, en una dualidad que ha sido llamada "el problema materia-símbolo".

El problema materia-símbolo probablemente tuvo su primera expresión en la Teoría de Autómatas Autorreplicantes de von Neumann (1966). En ella, se afirma que un sistema computacional autorreplicante consta necesariamente de dos partes: la parte física (hardware) y la programación (software). Von Neumann sugiere que una máquina que se dedique a replicar los elementos físicos no puede crear los símbolos necesarios para la parte de programación, y una que se dedique a producir el software no puede elaborar el hardware. Para una autorreplicación completa, son necesarios dos modos de producción complementarios.

Howard Pattee (1992) generaliza la visión de von Neumann y la aplica prácticamente a todo el universo físico, de una manera notablemente peirceana. Según Pattee, la dicotomía materia-símbolo puede encontrarse incluso en los problemas de física más sencillos, pues está relacionada con la diferenciación entre medición y ley. Las leyes físicas son abstracciones, independientes de las condiciones iniciales de los sistemas; las mediciones, por su parte, están relacionadas con las características físicas de un sistema en un momento dado. Ambas expresiones del mundo (leyes y mediciones) son dos facetas independientes pero complementarias de la realidad. Para Pattee, la medición del mundo físico y su expresión en leyes abstractas un proceso análogo a la percepción de objetos físicos por un organismo y su expresión en sistemas semióticos. En ambos casos, existe un

mediador que relaciona los dos fenómenos independientes: un científico que relaciona un conjunto de condiciones iniciales con una ley, o un organismo que relaciona una percepción del mundo con una imagen mental. El mediador toma el papel de interpretante en el contexto de la teoría semiótica de Peirce.

Pattee fue el primero que asoció esta dicotomía aparentemente universal con los sistemas biológicos. De hecho, afirma que el problema materia-símbolo se encuentra en el centro de las investigaciones acerca del origen de la vida. Los paralelos con las consideraciones de von Neumann son claras: el establecimiento de la primera célula requiere del surgimiento de un sistema físico, representado por el fenotipo, articulado con un sistema simbólico, contenido en el genotipo. Ambos se retroalimentan y se necesitan mutuamente para un funcionamiento continuo. El argumento fundamental de Pattee es que el modo simbólico del sistema logra escapar de las necesidades físicas, pues tiene un control autónomo sobre sus procesos. El sistema, aunque esté construído sobre elementos puramente físicos, con propiedades y características ineludibles, logra que a través de la interacción e interpretación de estos elementos se puedan lograr resultados distintos a pesar de que las condiciones del medio sean las mismas. Así, mientras que la parte puramente física del sistema obedece pasivamente las leyes físicas, la parte simbólica presenta mecanismos de autocontrol que permiten una variedad más grande de comportamientos.

Uno de los ejemplos de esta interacción entre materia e información es la distinción entre fenotipo y genotipo. Si rastreamos esta distinción a su origen encontraremos que éste se encuentra en la dualidad de las moléculas que componen el código genético, los aminoácidos y los nucleótidos. Estas moléculas existen en la realidad física y tienen propiedades particulares que permiten diferenciarlas unas de otras, pero el sistema que las usa (es decir, el aparato de traducción celular) las interpreta como símbolos relacionados entre sí: un codón *significa* un aminoácido, a través de una relación simbólica no determinada por leyes físicas.

Rocha (2001) elabora sobre la idea del corte epistémico de Pattee, afina la aplicación de la separación de materia-símbolo en sistemas biológicos y la lleva a la conclusión inevitable: si la dicotomía entre símbolo y materia está en la raíz de la separación de fenotipo y genotipo, entonces se encuentra también en la base de cualquier tipo de evolución abierta (“open ended evolution”), la cual es un principio fundamental de

la estructuración de cualquier tipo de vida conocida. Propone que existen tres requisitos para este fenómeno: 1) una base inerte (e incoherente) en la que se puede expresar el modo simbólico de la materia; 2) un conjunto de reglas semióticas para relacionarla con una función; y 3) un uso contextual del producto. Estos requisitos son una reinterpretación de la triada de Morris (1938, 1955) compuesta por sintáctica, semántica y pragmática respectivamente, que a su vez está explícitamente basada en las triadas peirceanas de semiosis. (Peirce, 1991).

Debido a la importancia de ambos modos de funcionamiento, en el presente trabajo se trata de tener siempre en mente la presencia de características físicas y simbólicas. En un modelo completo, o, en su defecto, una serie de modelos que representen satisfactoriamente los sistemas biológicos, tiene que haber una complementaridad entre los dos tipos de funcionamiento. No es posible desechar una mitad y pretender que la estructuración de los organismos se basa únicamente en características físicas, así como tampoco se puede depender de un modelo puramente simbólico. El modelo que nosotros presentamos se inclina, como se desarrolla más adelante, por analizar las características informacionales. Sin embargo, consideramos que es sólo parte de una modelación más compleja y completa, la cual deberá de tomar en cuenta de tres aspectos del código: el informacional, el fisicoquímico y la interacción de varios agentes para estructurar un lenguaje robusto. Este escenario se desarrolla con más detalle en la sección "Propuesta de estructuración histórica", en el capítulo 3.

Finalmente, aunque en general las propuestas de Pattee en las que insiste en ver a los seres vivos como sistemas fundamentalmente simbólicos (Pattee 1985, 2001) son una parte medular del presente trabajo, no podemos estar completamente de acuerdo con su visión (expuesta en Pattee, 1992) de considerar a todo el universo físico como un sistema semiótico. Una de las características fundamentales y más originales de Pattee es que provee de una base para diferenciar los sistemas vivos del resto del mundo físico. Afirmar que todo el universo, y no únicamente los seres vivos, hace uso constante de relaciones simbólicas elimina esta posibilidad.[2]

2 En sus últimos escritos, Pattee expresa opiniones similares. Por ejemplo, en Pattee (2001), comenta que "imagining such a subject-object distinction before life existed would be entirely gratuitous, and to limit control only to higher organisms would be arbitrary." Dichas frases hacen pensar que, aunque inicialmente trató de extender su visión dicotómica de símbolos a todo el universo físico, en las últimas etapas restringe la dinámica semiótica solamente a los seres vivos.

El problema significado - sentido

La dicotomía entre modo físico y modo simbólico, los distintos modelos y preguntas que cada uno implica y la complementaridad necesaria entre ambos, comienza a hacer evidente que hay dos polos posibles para acercarse al estudio de los sistemas semióticos ---naturales. Un sesgo completo a una visión materialista provoca que se trate de explicar a los sistemas semióticos únicamente por las propiedades de sus partes y por necesidades físicas. Así, se ignora por completo la interpretación que el sistema hace de sus elementos, y por lo tanto, la definición fundamental de signo de Peirce (1991): un signo es una cosa que representa otra cosa en los ojos del interpretante. De la misma manera, un sesgo hacia un estudio semiótico de los sistemas corre el riesgo de ignorar las características físicas de los elementos del sistema. En el caso extremo, este tipo de estudios consideraría a todas las relaciones existentes como completamente arbitrarias (Pattee, 2001; Barbieri, 2003, 2008), lo cual implica que un sistema es sólo una posibilidad entre muchas otras completamente equivalentes.

Entre ambos extremos existe un gradiente de posibilidades de estudio, y es en este rango en donde comienza a hacerse evidente el segundo gran problema. Existen algunos estudios, principalmente en el campo de la vida artificial o más recientemente en inteligencia artificial, que se inclinan hacia los modelos simbólicos pero no niegan la importancia de la cualidad arbitraria y universal de los símbolos. En ellos se considera, implícita o explícitamente, que las relaciones simbólicas son importantes como medio para llevar a cabo una función. Es decir, en el campo de vida artificial, una relación simbólica sólo se considera si produce una vía metabólica o alguna estructura ventajosa para el sistema. En los estudios de inteligencia artificial, se enfatizan las "imágenes mentales" que permitan nuevos comportamientos o percepciones útiles. Aunque todos los demás símbolos se consideran despreciables[3], deben estudiarse sin estas restricciones: existen símbolos que pueden establecer una relación arbitraria entre dos elementos sin que ésta represente una función inmediata o necesaria para el sistema. Pattee afirma que aunque un símbolo tiene

3 Incluso esta posición trae una serie de cuestionamientos de fondo, como señala Roy (2005): si la percepción fuera únicamente imágenes mentales, ¿quién es el que las interpreta? Así, una nueva capa de dificultad se agrega, tal vez necesariamente, a la interpretación de los símbolos desde la perspectiva de la inteligencia artificial.

una base material, no puede ser reducido a un estudio completamente materialista, ya que contiene algo más que su naturaleza física. Es posible ir un poco más lejos en esta argumentación, y afirmar que *un símbolo es algo más que una función asignada.*

Es así como se llega al segundo problema fundamental de esta discusión. Un símbolo puede tener una función dentro de un sistema biológico, evaluada en términos de ventaja, adecuación, novedad, etc. Si se considera el concepto "función" como equivalente al concepto "símbolo" (es decir, si se considera a la funcionalidad como característica necesaria y suficiente para definir a una relación simbólica), cualquier símbolo se reduce a la definición de Saussure: debe tener una imagen mental asociada, y ésta debe ocupar un lugar fijo en el contexto de un sistema semiótico determinado. Para aclarar un poco esta idea, se puede considerar la palabra "árbol", uno de los ejemplos clásicos de Saussure. Aunque la relación entre la palabra escrita o hablada y el concepto [árbol] es arbitraria, la palabra ocupa un lugar determinado en un lenguaje. La función de tal palabra será siempre evocar el concepto [árbol], y ha surgido en respuesta a la necesidad de nombrar ese objeto de alguna manera. Sin tal función, la existencia de la palabra no tendría sentido.

El escenario anterior es aplicado generalmente al origen de relaciones simbólicas en el lenguaje humano, pero existen otros tipos de sistemas semióticos en el mundo biológico. El código genético, que es el tema central del presente trabajo, aporta un claro ejemplo de esto. Las relaciones entre los tripletes de DNA y sus aminoácidos correspondientes son simbólicas: se trata de conexiones arbitrarias entre elementos de dos universos distintos (Barbieri, 2003). Además, no son funcionales, o al menos no inmediatamente. En su origen, la traducción de codones a aminoácidos no aportó de manera inherente una función al sistema: la función emerge una vez que una proteína específica es colocada en un contexto determinado.

De esta manera, podemos notar que existen dos formas para establecer un significado. En primer lugar, está el significado que tiene una función inherente, es decir, que puede ser entendido de inmediato en el contexto en el que surge. Un ejemplo de este tipo de "significado" es el que emerge en la traducción de una proteína, la cual en un contexto celular llevará a la síntesis de un nuevo metabolito, la formación de una molécula estructural o un proceso de comunicación o manufactura. Generalmente, los acercamientos a la biosemiótica analizan este tipo de significados: comunicación celular, endócrina,

inmunológica, expresión de funciones intracelulares y generación de rutas metabólicas. En segundo lugar, se encuentra el significado abstracto: un elemento *a* se relaciona simbólicamente con un elemento *b*. En dicho proceso de significación, no es posible considerar ni función ni comunicación. Ejemplos de ello son las traducciones de los diversos codones a aminoácidos, es decir, las relaciones que conforman el código genético. Las relaciones semióticas entre los elementos claramente están presentes, pero no poseen ese elemento semántico, necesario para establecer un proceso funcional: de cierta manera, se puede decir que son símbolos puros.

Umberto Eco ha diferenciado estos dos tipos de significados dividiendo a la semiótica en dos dominios: semiótica comunicativa y semiótica significativa, respectivamente. Hace notar que en ambos las relaciones simbólicas están presentes realizando su peculiar diagnóstico: "siempre que se manifiesta una posibilidad de mentir estamos frente a una función semiótica" (Eco, 1975). Cuando se considera que Eco llama a la parte comunicativa "teoría de producción de signos", y a la parte significativa "teoría de códigos", creo que resulta evidente cuál es el acercamiento que más conviene al análisis del origen, desarrollo y funcionamiento del código genético.

Una de las ambiciones de este trabajo es buscar los puntos en común entre ambos modos semióticos. Tanto conceptual como metodológicamente, existen muchas uniones entre la semiótica significativa y la semiótica comunicativa. El primer tipo es, evidentemente, el interés predominante, pues se analizarán aspectos del código genético dejando a un lado la parte semántica del genoma. Sin embargo, uno de los argumentos principales es que un código genético viable debe conservar una determinada configuración informacional para que los mensajes puedan ser comunicativos. Es decir, la traducción de una cadena en otra debe ser capaz de transmitir confiablemente un mensaje funcional. Así, se propone que en el origen del sistema las características informacionales y la funcionalidad del sistema pueden surgir de manera espontánea y son mutuamente dependientes.

2. Límites de los modelos

La semiótica puede abarcar cualquier tipo de objetos. Ya se ha discutido la visión de Peirce de un universo pansemiótico, en el que cada ser puede convertirse en un símbolo y representar otra cosa cualquiera. Sin llegar a tales extremos, es posible afirmar que la semiótica abarca mucho más de lo que inicialmente pudiera pensarse. Además de los símbolos, puede haber un acercamiento a un proceso, no sólo a través de símbolos, sino de *signos en general.* Es decir, es posible acercarse semióticamente a diversos procesos, físicos o mentales, mediante otras manifestaciones, como índices, iconos, funciones semánticas, etc. Por ejemplo, cuando una veleta indica la dirección del viento mediante un cambio en su posición, está ocasionando el surgimiento de un índice, un tipo de signo no simbólico, que se establece cuando hay una conexión física entre los dos objetos relacionados. En este caso, la conexión física es la fuerza del viento que mueve la veleta; otros índices, como las fotografías o los cambios de color en una sustancia, se producen por la acción de los fotones y las reacciones químicas, respectivamente.

A la luz de estas consideraciones, uno podría pensar que la biosemiótica proclama obviedades. Ante la multitud de signos (es decir, de manifestaciones semióticas) que existen en el universo físico, resulta inevitable cuestionarse si el manejo simbólico de información de los seres vivos es una característica especial. Por supuesto, existe una gran diferencia entre encontrar signos en la naturaleza y encontrar símbolos bien establecidos, lo cual es un evento inusual. Es todavía más raro encontrar sistemas simbólicos estructurados, articulados entre sí, para dar lugar a funciones complejas, autorreferencias y clausura semántica: estas particularidades se presentan exclusivamente en el manejo de la información que realizan agentes biológicos.

Aunque justamente este tipo de sistemas articulados, y en particular los códigos naturales, son los que competen al presente trabajo, es necesario establecer límites para ordenar el estudio y la argumentación. Esto se vuelve particularmente necesario al tratar de modelar el sistema, en cualquiera de las acepciones de la palabra “modelación”, puesto que se trata de un término que, como señala Goodman, es de los más promiscuos del vocabulario científico: puede referirse a “almost anything, from a naked blonde to a quadratic equation” (Goodman, 1972). Además, un modelo es necesariamente una

falsificación de la realidad (Turing, 1952), en la que se elige una serie de características que se suponen importantes y se deja fuera muchas otras que pueden introducir ruido a la dinámica de interés, o simplemente porque se piensa *a priori* que no tienen relevancia en el estudio. Además, hay una infinidad de variables y características que se incluyen o se excluyen por ignorancia de su existencia, características conceptuales propias del modelo, o incluso necesidades experimentales específicas. Tal inclusión y exclusión asimétrica es inevitable, e inseparable de la naturaleza de la modelación, por lo que se vuelve un paso necesario considerar al menos ciertos límites generales que enmarquen a la representación de la realidad en un modelo, y, por tanto, a los métodos utilizados y a las respuestas que se intentan obtener.

¿Qué tipo de modelo es el que presentamos en este trabajo? Reflexiono acerca de la sucesión de modelación propuesta por Eric Winsberg (1999), la cual se mueve desde la equivalencia del sistema a estudiar con un sistema físico simple ("este sistema se comporta *como si fuera* un oscilador armónico"), a través de una serie de ecuaciones diferenciales y su posterior discretización para ser resuelta, pasando por la eliminación o adición de factores e interacciones (modelación *ad hoc*) hasta el punto más complejo de la modelación: la modelación de un fenómeno, en la que se incluyen no sólo análisis matemáticos de los resultados de los modelos anteriores, sino también una integración con conocimiento teórico y empírico de diversos campos. La analogía física simple está presente en nuestro modelo; el análisis de Shannon de información es una cadena simple de Markov que presenta un modelo probabilìstico de cambios entre estados. También puede considerarse, definitivamente, un modelo *ad hoc*, pues dejamos fuera aspectos fundamentales del sistema real, como las caracterìsticas fisicoquìmicas de los elementos o el contenido semántico de la secuencia. Asimismo, intentamos conectar nuestros resultados con una estructura teòrica y experimental mayor. Por lo tanto, creo que es posible afirmar que nuestro modelo es un *modelo del fenómeno*, en el que nuestros resultados son sólo una parte pequeña de un análisis mayor que incluye evidencias evolutivas, fisicoquímicas y matemáticas y cuyo fin último es proponer un acercamiento limitado a un sistema extremadamente complejo. Una vez realizadas estas consideraciones superficiales acerca de cómo se quiere representar un fragmento del mundo físico, es necesario determinar la extensión de dicho fragmento, y qué escalas abarcará nuestro estudio.

Límites inferiores

Los límites inferiores de la investigación, y por lo tanto del modelo, son fundamentales. Cuando no se definen, no hay una manera clara de saber cuándo detenerse en el intento de representación de la realidad. Sin límites inferiores, existe el riesgo de tratar de modelar cada una de las partes del sistema y cada una de las interacciones entre ellas. Si se va más lejos, se puede comenzar a modelar las partes que componen a cada elemento del sistema, y seguir así *ad infinitum.* Los límites permiten evitar ese proceso de atomización, en el que cada una de las partes debe ser representada con todos sus componentes, y el cual conduce necesariamente a un reduccionismo extremo, en el que cualquier comportamiento o cualquier relación de causalidad se puede encontrar únicamente en las unidades más básicas de construcción. Este tipo de pensamiento se puede observar comúnmente en estudios que tratan de explicar particularidades etológicas o procesos ecosistémicos únicamente a través del estudio de las unidades más básicas de la vida, como los genes, sin considerar posibles influencias externas o comportamientos que emergen únicamente en niveles superiores, como el organísmico o el ecológico[4].

Existe otra consideración a favor de los límites en la modelación científica. Se podría pensar que el modelo ideal es el más apegado a la realidad, pero ocurre justamente lo opuesto. Consideremos por un momento un modelo que imite perfectamente cada punto del sistema de estudio, cada una de las interacciones entre sus componentes y las influencias externas a las que pueda estar sujeto. Tal tarea es imposible, por supuesto, pero en caso de que pudiera lograrse, se anularían por completo los usos y objetivos del modelo. El modelo perfecto sería una copia fiel del fenómeno físico de interés, y ambos (modelo y realidad) serían igualmente difíciles de estudiar.

En trabajos como el presente, establecer límites inferiores es un poco distinto a la mayoría de los estudios científicos. Si bien tratamos de estudiar entes físicos como aminoácidos y nucleótidos, el objeto de estudio no son precisamente las características

4 La integración de distintos niveles de descripción resulta una labor fundamental, pues, aunque sean fundamentalmente distintos, el flujo de información ocurre a través de ellos sin solución de continuidad (Farnsworth, Nelson y Gershenson, manuscrito sin publicar). Posiblemente un campo de investigación fértil se encuentre en el análisis de las regiones difusas de transición entre niveles, en los que hay simultáneamente cambios de características fundamentales y continuidad en varios procesos.

fisicoquímicas o las propiedades materiales de éstos; son las interacciones que ocurren entre ellos y el surgimiento de relaciones semióticas en el sistema. De hecho, una parte fundamental del planteamiento del modelo es abstraer los objetos entre los que se establecen las relaciones simbólicas, dejando de lado, al menos por ahora, cualquier particularidad física. Es cierto que, en los sistemas simbólicos naturales, las leyes físicas determinan gran parte de la dinámica (Pattee 1995, 2001), pero este trabajo trata de explorar las características puramente informacionales y semióticas de los organismos. Por lo tanto, es necesario establecer niveles de complejidad distintos que en los estudios físicos. Usualmente, los niveles de organización se derivan de la aparición de propiedades emergentes: un ser vivo se puede estudiar desde el nivel molecular, celular, tisular, organísmico o ecológico. En un sistema biológico también es posible discernir diversos niveles de organización semiótica, y es en esta escala en donde se debe elegir cuáles de estos niveles incumben a la investigación realizada, y cuales son demasiado atomizados y complejos.

Según Umberto Eco (1968), el nivel básico de la semiótica es el que se compone únicamente de estímulos y señales. Él afirma que la semiótica no debe confundirse con la semántica, disciplina que "se ocupa (o finge ocuparse) del 'sentido'" (pp. 29). En cambio, la semiótica también se ocupa de sistemas de significación, pero de los cuales no necesariamente ha emergido un mensaje determinado. A partir de este nivel básico, la semiosis tiene el potencial de ser más compleja. Las señales se pueden traducir utilizando un código, es decir, comunicarse a través de un canal para detectar errores y ser utilizadas en otro sistema. Las señales traducidas se pueden usar para formar mensajes, dándole contendo semántico a dichas secuencias, y éstas a su vez pueden articularse en un sistema completo que tenga diversos tipos de autorreferencia, la cual se expresa en los seres vivos como la capacidad de reproducción. En los sistemas biológicos, la semiosis abarca muchos niveles más: a pesar de que las células funcionan como unidades delimitadas, comunican mensajes a otras células de diversos tipos a nivel incluso organísmico. Además, los organismos tienen sistemas semióticos para comunicarse con miembros de su misma especie o de especies distintas, hasta llegar a la comunicación cultural, cuya complejidad es mayúscula. Por supuesto, este recuento de los diferentes niveles de semiosis en los seres vivos (señales --> códigos --> semántica --> comunicación intracelular -->comunicación

intercelular --> comunicación entre organismos --> comunicación cultural) es sólo un boceto superficial. Hay múltiples estados intermedios de transición y matices entre cada uno ellos.

Según las consideraciones anteriores, el nivel inferior de la semiótica está marcado por la comunicación lineal de señales, y por tanto es posible comenzar a determinar los límites de la presente investigación considerando los alcances teóricos de dicho nivel. El campo de estudio que se ocupa de este tipo de comunicación es la teoría de la información, cuya actividad central es medir qué tanta información está contenida en un mensaje. Los tres puntos fundamentales de este análisis son la fuente del mensaje, el canal a través del cual se transmite y el receptor. Además, la teoría de la información considera los cambios que sufre el mensaje al entrar y salir del canal, así como los posibles efectos externos que pueden ocasionar errores en la transmisión. Según Robert Ash (1990) la teoría de la información se resume en el estudio del siguiente teorema: "es posible transmitir información a través de un canal ruidoso en cualquier tasa por debajo de la capacidad de dicho canal, con una probabilidad de error arbitrariamente pequeña". La estructura general de los componentes de este sistema se muestra en el siguiente esquema:

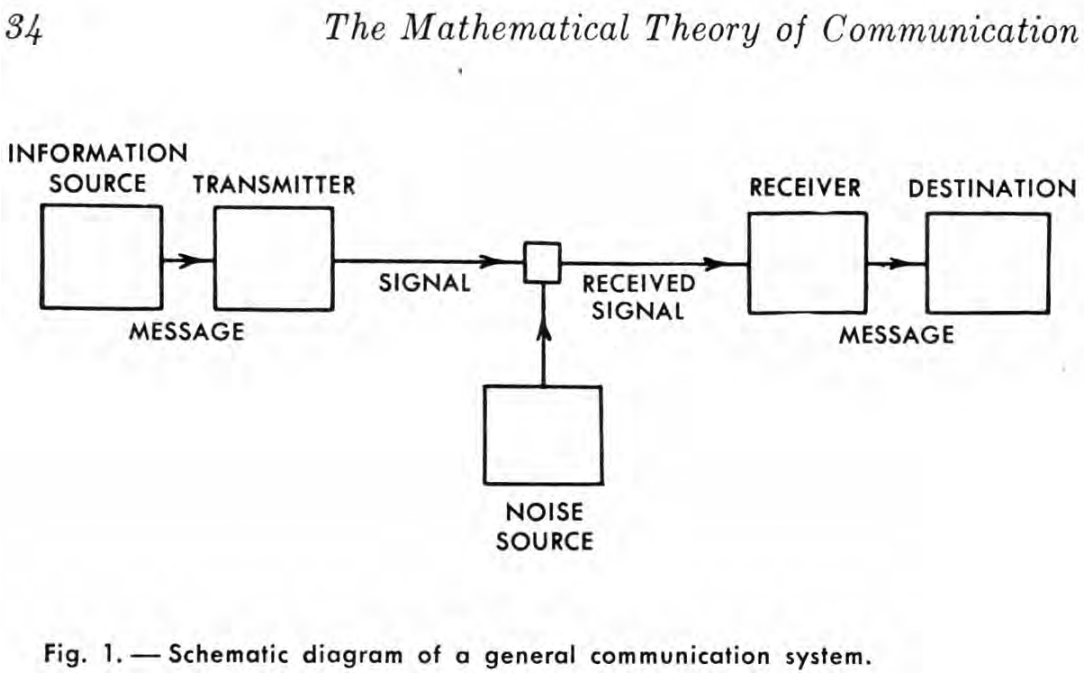

Fig. 1: El esquema original de Shannon, 1948

La teoría de la información tiene su origen en un artículo clásico de Claude Shannon (1948), en el que se propone un método de análisis matemático cuyo interés inicial era la transmisión de mensajes telegráficos. En resumen, trata de determinar la posibilidad de comunicación literal de un mensaje; es decir, la transmisión de un mensaje de una fuente a un receptor sin ningún cambio. Esto, por extensión, incluye los límites de compresión de información y su paso a través de un canal con un grado de ruido variable. En el centro de la teoría se encuentra el concepto de entropía informacional, que se expresa como una medida probabilística de poder adivinar las variables aleatorias que representan cada una de las señales del mensaje. En cierta manera, la entropía informacional es una medida de la incertidumbre, determinada por los caracteres que ya han aparecido en el mensaje y modificada por los caracteres nuevos que sean recibidos. Si alguna letra, por ejemplo A, aparece muchas veces en un mensaje, la entropía informacional de éste será pequeña, y tendremos una mayor certidumbre que la siguiente señal que recibiremos será la letra A.

A pesar de que a primera vista pueda parecer limitada, la teoría de Shannon es inherentemente tan vasta que desde su publicación ha encontrado una enorme cantidad de aplicaciones. Se ha utilizado en aplicaciones tecnológicas de transmisión y almacenaje de información, así como compresión de mensajes (por ejemplo, archivos informáticos), como es obvio; además se puede aplicar a cualquier proceso que conlleve incertidumbre para analizarlo matemáticamente. Por ejemplo, en las ciencias biológicas puede ser utilizada para investigar características informacionales de las secuencias de DNA o aminoácidos, e incluso se utiliza como una medida de biodiversidad: si se considera a cada tipo de organismo como un "caracter", y se estudia la frecuencia de encuentros que se tiene con cada uno, es posible inferir qué estructura tiene la comunidad biológica estudiada. Si un ecosistema arroja un resultado de entropía informacional alto, es probable que haya un número similar de cada tipo de organismos; si la entropía de Shannon es baja, probablemente en el ecosistema esté sobrerrepresentado un grupo reducido de especies. En cierta forma, esto podría considerarse como una manera en que el universo es pansemiótico, como afirmaba Peirce, por lo que se puede representar a cualquier cosa, incluídos los seres vivos, como signos y, de manera potencial, símbolos.

Podría pensarse que la teoría de la información es ideal para los fines del presente trabajo, ya que se compone de una serie de conceptos y metodologías que describen de

manera matemáticamente precisa una serie de características de un mensaje. Sin embargo, hay una gran limitante para los alcances de la teoría de Shannon y de las medidas derivadas. Se mencionó un par de párrafos atrás que la teoría de la información tiene por objeto el estudio de la comunicación *literal* de un mensaje. Esto quiere decir que dicho mensaje se comunica de manera exacta, sin ser interpretado ni transformado en otro distinto. Tal es uno de los puntos centrales de la propuesta de Shannon: sin interpretación ni traducción, el mensaje no tiene contenido ni función, sino que se estudia como una simple concatenación de señales, y se descarta desde el inicio todo el estudio de contenido semántico. En palabras de Shannon (1948): "El problema fundamental de la comunicación es reproducir, exacta o aproximadamente, un mensaje que se ha elegido en otro punto. Frecuentemente los mensajes tienen significado; es decir, se refieren o están correlacionados con una entidad física o conceptual, según algún sistema. Dichos aspectos semánticos de la comunicación son irrelevantes al problema de ingeniería. El aspecto significativo es que el mensaje será elegido de un conjunto de mensajes posibles". Según Battail (2009), en esta limitación autoimpuesta reside toda la fuerza de la teoría de la información: al no preocuparse por interpretaciones, funciones o incluso por códigos semánticos, el mensaje puede ser analizado por parámetros estrictamente cuantitativos, sin ninguna clase de ambigüedad.

Es cierto que ignorar la semántica confiere ciertas ventajas. El análisis se vuelve mucho más simple al no tener que lidiar con los problemas semánticos inherentes a cualquier lenguaje, los cuales han señalado, entre muchos otros, Wittgenstein (1988) y Gödel (1992). Además, al ser completamente matemático, es posible aplicar el conocimiento obtenido acerca del mensaje al diseño de algún aparato de comunicación, sabiendo precisamente el grado de fidelidad máxima y los pasos necesarios para acercarnos a él. Sin embargo, la falta de análisis semántico o cualquier otro tipo de acercamiento semiótico es la razón por la cual la teoría de Shannon resulta insuficiente para los fines de este trabajo. El objeto de esta investigación es relacionar las características informacionales de un mensaje con significado funcional con los códigos utilizados para realizar una traducción viable, cuyo producto posea características similares o análogas al mensaje original, conservando su funcionalidad. Así, la teoría de Shannon resulta un marco conceptual inapropiado, pues ignora voluntariamente dos procesos de importancia

fundamental para nuestros fines. Por un lado, la comunicación no literal, es decir, la producción de un mensaje distinto al original a través de un proceso de traducción por un código. Por el otro, el contenido semántico del mensaje, es decir, la función que dicho mensaje llevará a cabo en el contexto celular en el cual se ha producido. Por lo tanto, es necesaria una serie de metodologías y conceptos que pueden tener como cimiento la teoría de la información, pero la expandan permitiendo la inclusión de significados.

Límites superiores

Así como Umberto Eco trata de determinar los lìmites inferiores en *La Estructura Ausente* (Eco, 1968) y posterioremente en su *Tratado de Semiótica General* (Eco, 1975), también expone lìmites superiores de la semiótica como rama del conocimiento. Aunque tomo como base su delimitación, sus argumentos y objetivos son un poco distintos de los de la presente investigación. Por un lado, Eco afirma que en el límite inferior se debe de considerar los estímulos y señales informativos, pero en definitiva los fenómenos genéticos y neurofisiológicos no son de competencia para la semiótica. Por otro lado, Eco propone que el estrato superior esta disciplina incluye la cultura, la cual se puede definir por tres eventos identificables: el uso de un objeto (S) para una función (F), la determinación de la relación entre objeto y función, y la capacidad de reconocer una serie de objetos comparables (S1, S2, S3, S4... Sn) de acuerdo con un tipo abstracto (P). Aunque el proceso de asociaciones abstractas esté enfocado a realizar una comunicación entre individuos, estos tres aspectos de la cultura no dependen de la presencia de la comunicación como tal. Esto quiere decir que aunque el individuo esté aislado, es suficiente la *posibilidad de comunicación* para que pueda emerger un comportamiento cultural.

La biosemiótica no tiene que seguir el mismo esquema de la semiótica general. Sin embargo, creo que la semiótica biológica es un campo lo suficientemente complejo y variado, en el que existen las diferencias suficientes como para que sea un estudio completo en sí mismo, y no únicamente un nivel o una pequeña parte de la semiótica. Ya se ha hablado de los límites inferiores de la biosemiótica en la sección anterior; también se ha dicho que coinciden casi exactamente con los de nuestro modelo. En el caso de los límites superiores, la correspondencia entre la semiótica, la biosemiótica y este estudio es menos

clara: el mismo Umberto Eco no demarca un límite superior, sino que argumenta que la comunicación cultural debe de ser estudiada por la semiótica. Debido a que el presente trabajo es una investigación acerca del origen de un sistema semiótico simple, el aún difuso límite superior no necesita ser explorado; además, la complejidad de un sistema con comunicación cultural hace que su representación en un modelo sea extremadamente difícil.

Un lector atento notará que no trato de realizar ningún intento de buscar equivalencias entre los límites superiores de la semiótica, la biosemiótica y el presente trabajo. Puesto que los estudios semióticos traducionales establecen como un difuso límite superior a "la cultura humana", necesariamente manejan convenciones antropocéntricas. La biosemiótica, por su parte, posiblemente tenga que hacer uso de la visión casi cósmica de Peirce: no depende de convenciones humanas sino de un interpretante cuya naturaleza es variable. Además existe una diferencia importante en sus límites: en su límite inferior se pueden atomizar los componentes y proponer un punto específico en el que el fenómeno semiótico desaparece; en el límite superior, es no es facil (y tal vez ni siquiera sea posible) realizar una operación análoga y tratar de diagnosticar una desaparición de la semiótica por aumento en la escala de descripción. Por otro lado, para tratar de entender los alcances de nuestro modelo, necesitamos dibujar límites propios a la investigación, aunque éstos sean completamente artificiales.

La definición de nuestro sistema de estudio es, pues, fundamental para determinar qué rango de los procesos semióticos se intenta modelar, y qué tipo de resultados esperamos obtener. Algunas características del sistema de estudio son las siguientes:

- La semiosis que queremos estudiar tiene que estar basada en un código unívoco. Es decir, a diferencia de la semiótica Peirceana, debe excluir procedimientos hermenéuticos y funcionar exclusivamente por relaciones pareadas entre dos alfabetos de signos para evitar ambigüedad.
- Debe ser capaz de surgir espontáneamente a través de un proceso de autoorganización, lo cual implica necesariamente que no hay un mecanismo de control central, sino que emerge de la interacción de los diversos elementos que la componen.

- Es necesario que exista algún reservorio de las relaciones semióticas creadas, para que el código pueda permanecer sin disiparse a través de un lapso significativo de tiempo. En la mayoría de los casos de los sistemas biológicos, esto ocurre cuando hay adaptadores que pueden ser codificados en el genoma y ser expresados continuamente. En el caso del origen del código genético, que necesariamente antecede a la emergencia de mensajes codificantes, la permanencia de las relaciones simbólicas probablemente reside en las relaciones fisicoquímicas entre los elementos y otros mecanismos, los cuales decidimos no modelar.
- Debe cumplir con la dualidad propuesta por Pattee (1995, 2001): ser arbitrario y, simultáneamente, estar sujeto a características fisicoquímicas. Si se toma en cuenta estas últimas para facilitar el análisis y la modelación, es conveniente que se consideren únicamente las interacciones simples entre dos elementos, como los mecanismos de reacción propuestos por Copley *et al* (2005).
- El sistema debe ser lo más sencillo posible, pero conservando las características neesarias de los sistemas simbólicos: semiosis (preferentemente en la forma de un código), información, posibilidad de surgimiento de mensajes funcionales. Es decir que para un acercamiento inicial de una investigación acerca del surgimiento de la semiosis biológica, el sistema de estudio debe ubicarse en el límite inferior de la semiótica.

Así, a la pregunta "¿es posible marcar límites a los fenómenos semióticos?" propongo una respuesta torpe: creo que la semiótica y la biosemiótica son sólo dos partes de una semiótica universal que entrevió Charles Peirce. Los límites superiores de la primera (la cultura humana) no corresponden a los de la segunda (¿los ecosistemas?), y ninguna de las dos facetas agota las posibilidades de los fenómenos semióticos posibles. Afortunadamente, el presente trabajo es un modelo restringido, y no necesita más que sus límites propios, autoimpuestos y seguramente artificiales para observar a su objeto de estudio. Como se argumenta en la sección de Discusiones, aún un modelo tan limitado como este puede traer consecuencias argumentativas que se extienden tanto espacial como temporalmente fuera de las fronteras de la modelación.

3. El código genético

Dicotomías

Hasta hace unas décadas, los sistemas simbólicos eran exclusivamente arbitrarios. Prácticamente todas las investigaciones se enfocaban a sistemas en las que un símbolo era creado por convenciones humanas. Por ejemplo, gran parte de los trabajos trataban de rastrear el surgimiento y las relaciones de las palabras en distintos idiomas que se refieren a un mismo concepto. Otros trataban de determinar qué tan concretos eran los significados de las palabras o las asociaciones simbólicas (religiosas, emblemáticas, psicológicas) de un concepto con otro. Algunos, tal vez más profundos, intentaban saber si los símbolos mentales son simplemente construcciones artificiales, en cierto sentido parasíticas a nuestra mente, o si tienen una representación en algún nivel de la realidad externa.

La concepción de los sistemas simbólicos y sus posibilidades cambió radicalmente con el origen y el desarrollo de la biosemiótica. Esta rama de la biología mostró por primera vez símbolos que no existían exclusivamente en la mente humana, sino que se componían por elementos físicos y tangibles. Esta dicotomía implica que los sistemas naturales de codificación poseen tanto elementos arbitrarios, derivados de su naturaleza semiótica, como necesidades físicas, derivados de su composición por partes tangibles como moléculas y células, y no únicamente por abstracciones.

La oposición entre los elementos físicos y abstractos de los códigos naturales tuvo su expresión más elocuente en los trabajos de Howard H. Pattee. En uno de sus argumentos más expansivos, centrado en el código genético, él afirma que existe una discontinuidad necesaria entre sujeto y objeto, o más precisamente, entre el objeto y su representación simbólica creada por el sujeto. Más aún, propone que, contrariamente a lo que se piensa, esta distinción no sólo existe en la mente de los sujetos con un cerebro altamente organizado, sino que está presente en cualquier comportamiento simbólico. Llama a esta discontinuidad un "corte epistémico", que sólo puede ser resuelto cuando el sistema sujeto hace una serie de distinciones entre diversos estados del objeto, y estas distinciones no

están sujetas a una ley física. Es decir, que para que un sistema funcione de manera simbólica, sea cual sea el funcionamiento específico, debe de existir una serie de "reglas", las cuales no se derivan de manera determinista de ninguna ley física (es decir, son arbitrarias), *aún estando expresadas en sistemas físicos que deben de obedecer dichas leyes*. En ese sentido, el genotipo y el fenotipo están conectados por una relación simbólica arbitraria, con una serie de restricciones flexibles, que en el momento de funcionar en un contexto celular originan toda la dinámica del proceso de la vida. Quiero hacer énfasis en la necesidad del contexto, pues fuera de él el funcionamiento simbólico no tendría sentido alguno. Como afirma Pattee: "a single folded protein has no function unless it is a component of a larger unit that maintains its individuality by means of a genetic memory". (Pattee, 2001). La individualidad de la que habla Pattee sólo se puede lograr si ambos lados del corte epistémico están coordinadas para provocar la continuación de la existencia de su contexto, y el contexto permite que dicho corte epistémico se exprese funcionalmente. Es decir, el sistema es autorreferente, o, en palabras de Pattee, existe clausura semántica ("*semantic closure*") (Pattee, 1982, 1995). ; esta autorreferencia se encuentra en la médula del problema del origen de la vida, y su consecuencia más conocida es el llamado "problema del huevo y la gallina".

Los elementos puramente simbólicos de los códigos naturales, en general, y del código genético, en particular, es decir, las relaciones de correspondencia entre un universo de símbolos y otro, permiten ver de manera característica cómo incluso las abstracciones deben de tener una base material. En los códigos naturales, las correspondencias simbólicas se realizan a través de *adaptadores físicos*. Por ejemplo, en el código genético los símbolos dependen de la presencia de tRNAs que especifican la secuencia del codón que se traducirá, y de la acción de aminoacil-tRNA sintetasas que unen al tRNA con el aminoácido que les corresponde. La manifestación de la abstracción simbólica en el plano puramente físico también se puede observar en el uso que el sistema biológico le da a la traducción: una serie de reglas abstractas sirven para construir elementos físicos, que a su vez llevan a cabo funciones específicas dentro de la célula.

Esta irrupción directa de los elementos abstractos en la realidad física de un sistema biológico implica que los símbolos están sujetos a un proceso de selección. De hecho, la

selección que actúa elementos simbólicos es una de las bases conceptuales de este trabajo. Nosotros suponemos que varios factores relacionados con el funcionamiento semiótico e informacional pueden ser optimizados, en mayor o menor grado, al menos en etapas tempranas de su formación. Además, suponemos que debido a que los organismos son entes sujetos a su historia, el código genético también lo está (siguiendo a Knight et al, 1999); este factor temporal implica indirectamente que la estructura organizada se debe a un proceso de interacción constante de sus partes a través de miles de años, hasta alcanzar el estado estable que conocemos actualmente. Es indispensable tomar en cuenta que este proceso de estructuración no estuvo dirigido por un diseñador central ni contó con un objetivo *a priori*, por lo que los patrones actuales surgieron por autoorganización. Una última suposición conceptual de este trabajo es que el código genético no representa un estado inmejorable de eficiencia. Actualmente, se conoce un sólo código genético; las variantes de algunos organelos u organismos unicelulares son modificaciones menores y, muy probablemente, relativamente recientes. Sin embargo, aunque la configuración de relaciones entre codones y aminoácidos es universal, es sólo un estado de los muchos posibles. En este punto seguimos la hipótesis de Francis Crick: en cuanto un código se ha establecido con cierto grado de eficiencia y los efectos negativos del cambio de un codón se vuelven insostenibles, la evolución de éste se detiene. Osawa y Jukes (1992) propone un número de escenarios en los que los codones pueden ser reasignados. El proceso más factible es la desaparición casi total de un codón en el dialecto genético de un organismo, junto con un cuello de botella en la población de los individuos de la especie.

La aparente inmutabilidad del código genético estándar no implica que éste sea la única opción factible o viable. El conjunto de relaciones simbólicas puede tomar varios estados posibles, muchos de los cuales son igual o más eficientes que el código estándar. Esta afirmación es una de las bases conceptuales de la presente investigación. Dentro del conjunto de códigos posibles, los distintos procesos que dieron origen y estructura al presente código genético delimitan un subconjunto de códigos viables que cumplan las necesidades fisicoquímicas, informacionales y de corrección de errores. Los elementos de este subconjunto son más o menos eficientes que el código estándar, pero comparten algunas de sus características fundamentales. Para tratar de precisar cuáles son estas características, es conveniente analizar el código genético estándar y sus patrones.

Optimización

La optimización en sistemas biológicos, y en particular en procesos evolutivos, acarrea una serie de controversias inevitables. En general, se relaciona inmediatamente con mecanismos teleológicos: la optimización es considerada un proceso dirigido a un fin determinado. Como tal, es un proceso sin cabida en la estructura de la teoría de la evolución actual, que está dominada por el mecanismo de selección natural, la cual a su vez se basa en el surgmiento de variaciones al azar.

Nosotros argumentamos que la optimización, contrario a lo que usualmente se piensa, no tiene un componente inherentemente teleológico. La optimización es un problema de búsqueda de máximos o mínimos (Haefner, 1997), que puede ser llevada a cabo por exploración azarosa de un paisaje adaptativo. Por supuesto, este tipo de mecanismo no garantiza la llegada a un máximo global, e incluso ocasiona que la mayoría de las veces una población determinada se establezca indefinidamente en un máximo local con adecuación relativamente baja. La mayoría de las veces, este estancamiento en un máximo local puede ser revertido únicamente por el surgimiento de una mutación radicalmente distinta al promedio de la población, o por la modificación del paisaje adaptativo. Por supuesto, tanto la topografía del paisaje adaptativo como la presencia de una población en distintos puntos de él son simplificaciones y abstracciones de lo que ocurre en el mundo físico.

Existen varios procesos en el mundo biológico que presentan signos de optimización. Aunque su eficiencia no es máxima, el uso y la obtención de energía en los seres vivos se realiza de tal manera que muchos autores lo han considerado una estrategia de disipación termodinámica que genera autoorganización (Brooks y Wiley, 1986; Nicolis y Prigogine, 1989; Schneider y Kay, 1994; Prigogine, 1997); el sistema de información, es decir, el genoma de los organismos, tiene una serie de puntos de detección y corrección de posibles errores en la expresión. El metabolismo y los procesos celulares cuentan con *checkpoints* que deben de ser cubiertos satisfactoriamente para continuar su camino, como es el caso de la división celular; incluso a niveles mayores, las redes ecológicas se han estabilizado en configuraciones robustas, en las que en la mayoría de los casos se observa una

reconfiguración que evita el colapso total de la comunidad. El objeto de estudio de esta investigación (es decir, el sistema de correspondencias simbólicas que constituyen el código genético estándar) también presenta una serie de características que indican una optimización en varios aspectos de su funcionamiento.

Cuando existe optimización, no necesariamente hay una configuración única que sea la mejor. Es posible diseñar códigos con distintos grados de eficiencia para diferentes parámetros. Por ejemplo, cuando el énfasis se encuentra en la capacidad de detección de errores, es necesario determinar prefijos en el código que indiquen errores de traducción o de transcripción del mensaje original. Por ejemplo, el código postal estadounidense llamado ZIP+4, explicado por Yockey (2005) es altamente eficiente para detectar errores. A cada número del 0 al 9 le es asignado arbitrariamente una cifra de 5 dígitos binarios. Un cambio de un dígito en algún paso de la transmisión de información origina una cadena de 5 dígitos que no está asociada a ningún número del 0 al 9; es decir, esa cadena errónea es *non-sense* (no tiene sentido). La figura 2 muestra las 10 cadenas de dígitos asignadas, y debajo de ellas, las 5 cadenas *non-sense* posibles. Cuando el sistema detecta una cadena sin sentido, es evidente que ha cometido un error y es necesario rectificar la información. Este tipo de detección es imposible en el código genético, pues está saturado de asignaciones, y no hay suficientes codones *non-sense* como para establecer un sistema similar.

0	1	2	3	4	5	6	7	8	9
11000	00011	00101	00110	01001	01010	01100	10001	10010	10100
01000	10011	10101	10110	11001	11010	11100	00001	00010	00100
10000	01011	01101	01110	00001	00010	00100	11001	11010	11100
11100	00111	00001	00001	01101	01110	01000	10101	10110	10000
11010	00001	00111	00100	01011	01000	01110	10011	10000	10110
11001	00010	00100	00111	01000	01011	01101	10000	10001	10101

Fig. 2: El código postal ZIP + 4 (tomado de Yockey 2005)

También es posible enfocarse en la reducción de los errores, en cuyo caso el código estaría diseñado para que la mayoría de las mutaciones sean silenciosas mediante el uso extensivo de la redundancia. Estos códigos se pueden hacer aún más eficientes mediante varios

métodos computacionales, para explorar diversas posibilidades cercanas, pero incluso estos métodos son completamente deterministas: dependen en gran parte de la configuración inicial del código y del método de exploración. Es muy probable que el grado más alto de eficiencia que alcancen se encuentre en un máximo local. Esto quiere decir que aún teniendo un objetivo particular, un estado inicial diseñado específicamente y un método para alcanzar una eficiencia máxima, el sistema no necesariamente llegará a un máximo global. Por supuesto, el código genético no fue diseñado, y su evolución no tuvo ningún componente teleológico; además, las características físicas de cada elemento y las contingencias inherentes a cualquier proceso físico hacen aún menos determinista su proceso de optimización. El código genético estándar, que representa la culminación de este proceso, reduce enormemente los efectos de los errores de traducción, pero no se encuentra en el máximo global de eficiencia en ningún sentido.

La aparente optimización del código genético presenta uno de los más grandes problemas conceptuales en las propuestas de su origen y desarrollo. Como Freeland y Hurst (1998) han argumentado, la estructura que permite la corrección de errores tan eficiente que su surgimiento por procesos completamente aleatorios es altamente improbable. Sin embargo, la importancia central del código en la expresión de moléculas funcionales de los sistemas biológicos hacen que su modificación sea extremadamente difícil: al asignar a un codón un aminoácido distinto, se ocasiona que todas las proteínas del genoma del organismo mutante sean modificadas de manera extensiva, y la probabilidad de perder funcionalidad es alta. Tales consideraciones han llevado al surgimiento de teorías como la del accidente congelado (Crick, 1968; criticada recientemente en Sella y Ardell, 2006, y Söll y Rajbandary, 2006), que propone que después de un periodo de evolución temprana, se hizo imposible cambiar las relaciones simbólicas de la traducción sin efectos negativos o incluso letales para los organismos en cuestión. La teoría del accidente congelado y las ideas que de ella se derivan tienen implicaciones importantes. La más importante es que el código genético actual es un estado eficiente mas no óptimo, en el que es posible ver una serie de artefactos cuya presencia supone una reducción de las capacidades del código. Uno de los más evidentes es el aminoácido arginina y su asociación simbólica a seis codones (AGR y CGN), que, como se explica en la siguiente sección, es una constante excepción a los patrones del código, y posiblemente se haya asignado a través de una asociación

estereoquímica. Por lo tanto, es de importancia fundamental entender los patrones de diversos tipos que existen en el código genético estándar.

Regularidades y patrones

El código genético no es simplemente una colección de reglas dispuestas de manera aleatoria. Cuando se ordenan los aminoácidos en familias de codones, empieza a ser evidente que ciertas propiedades se correlacionan con las distintas posiciones de los nucleótidos de los codones. Por ejemplo, la segunda letra de los codones (usualmente dispuesta como las columnas del código genético) está correlacionada con las propiedades fisicoquímicas de los aminoácidos codificados. Esta es una parte fundamental del código; lo más probable es que hable tanto de la historia evolutiva temprana del sistema como de una posible tendencia a reducir errores en la traducción. Se pueden identificar dos tipos de regularidades en el código: fisicoquímicas e informacionales.

Nosotros no esperamos que el trabajo reproduzca al pie de la letra la estructura del código genético: para ello, tendríamos que incluir datos fisicoquímicos en el modelo computacional, y elaborar reglas de asociación que muy probablemente serían arbitrarias. No obstante, es necesario conocer este "código dentro del código" (Taylor y Coates, 1989), tanto para entender el escenario hipotético que se propone en la siguiente sección como para saber qué elementos del código genético intentamos modelar (las informacionales) y cuáles dejamos fuera de manera consciente (las fisicoquímicas).

Características fisicoquímicas.

En el código genético, las familias de codones similares codifican para aminoácidos similares. Cada una de las tres "letras" que componen los codones (es decir, cada uno de los tres nucleótidos que se interpretan dentro de la célula como un aminoácido en el momento de la traducción) tiene una asociación particular con alguna característica de los aminoácidos. Estas relaciones no fueron evidentes en los primeros años de estudio del código genético, y es posible que Wong (1975) sea el primer autor que publicó explícitamente la relación entre los codones y la fisicoquímica de cada aminoácido. Posteriormente Taylor y Coates (1989) desarrollaron y expandieron esa idea.

La primera letra de un codón, por ejemplo, se relaciona directamente con las rutas biosintéticas de los aminoácidos. Cada uno de los nucleótidos está relacionado con una serie de aminoácidos que comparten precursores metabólicos. Por ejemplo, los codones iniciados con A son la familia del aspartato, que incluye a la asparagina, treonina, isoleucina, metionina y lisina; el aspartato no está incluído dentro de esta familia sino que es el precursor de cinco aminoácidos mencionados. Los iniciados con U son casi todos aromáticos, y se agrupan en la familia del shikimato, el cual se deriva a su vez del fosfoenolpiruvato (PEP) y la eritrosa 4 fosfato (E4P). Los aminoácidos sintetizados por esta ruta son triptofano, fenilalanina y tirosina. Otros aminoácidos codificados por los codones que inician en U (cisteína y serina) se relacionan directamente con el triptofano mediante su ruta biosintética. Los codones que comienzan en C son la familia del glutamato, que incluye arginina, glutamina y prolina. De manera análoga al caso del aspartato, el glutamato no está incluído dentro de su familia, sino que es el únicamente el precursor. Finalmente, los codones que inician con G son un poco más ambiguos; no tienen ninguna ruta biosintética ni precursores comunes. Sin embargo, existe una cierta regularidad: todos sus aminoácidos (glicina, alanina, valina, aspartato y glutamato) son relativamente sencillos, y son los que se encuentran en concentraciones mayores en los experimentos de química abiótica, lo cual quiere decir que su síntesis es fácil de realizar. Además, los aminoácidos cuyos codones comienzan en G se sintetizan con una sola reacción química de sus precursores (ya sean éstos parte de la glucólisis o del ciclo de ácidos tricarboxílicos).

Como señalan Copley et al (2005), en los primeros pasos de las rutas biosintéticas comunes se encuentran compuestos derivados de las rutas metabólicas centrales (y posiblemente originales, según la propuesta de Smith y Morowitz, 2004) de la vida, ya sea el ciclo de Krebs o la ruta de glucólisis. Los codones iniciados en U son los aminoácidos aromáticos, sintetizados por la llamada ruta del shikimato, compuesto formado por la unión química de fosfoenolpiruvato y eritrosa-4-fosfato. Los codones que inician en A son la familia de aminoácidos sintetizados utilizando el aspartato, y más atrás en la ruta bioquímica, todos tienen la raíz en el oxaloacetato. Los codones iniciados con C codifican a la familia del glutamato, cuyo precursor es el alfa-cetoglutarato. Los que inician con G son más heterogéneos, pero comparten la característica de que se sintetizan a partir de una transformación simple de un compuesto de las rutas metabólicas antes mencionadas; la

relativa sencillez de síntesis explica que sean los que se forman en más abundancia en los experimentos de química prebiótica. En la siguiente ilustración (Fig. 3) se detallan las rutas de biosíntesis de los aminoácidos, y se puede ver claramente los agrupamientos con base en la primera letra del cdón: en la parte inferior, a partir del α-cetoglutarato (α-KG) se observa la mayoría de los codones CNN. Inmediatamente arriba, los codones ANN derivados de oxaloacetato (OAA); después, los codones UNN que se sintetizan por la ruta del shikimato y sus dos precursores. Por último, los codones GNN no se agrupan en filas, sino en una columna del lado derecho del esquema, pues se sintetizan a través de unúnico paso: una aminación simple de diversos precursores.

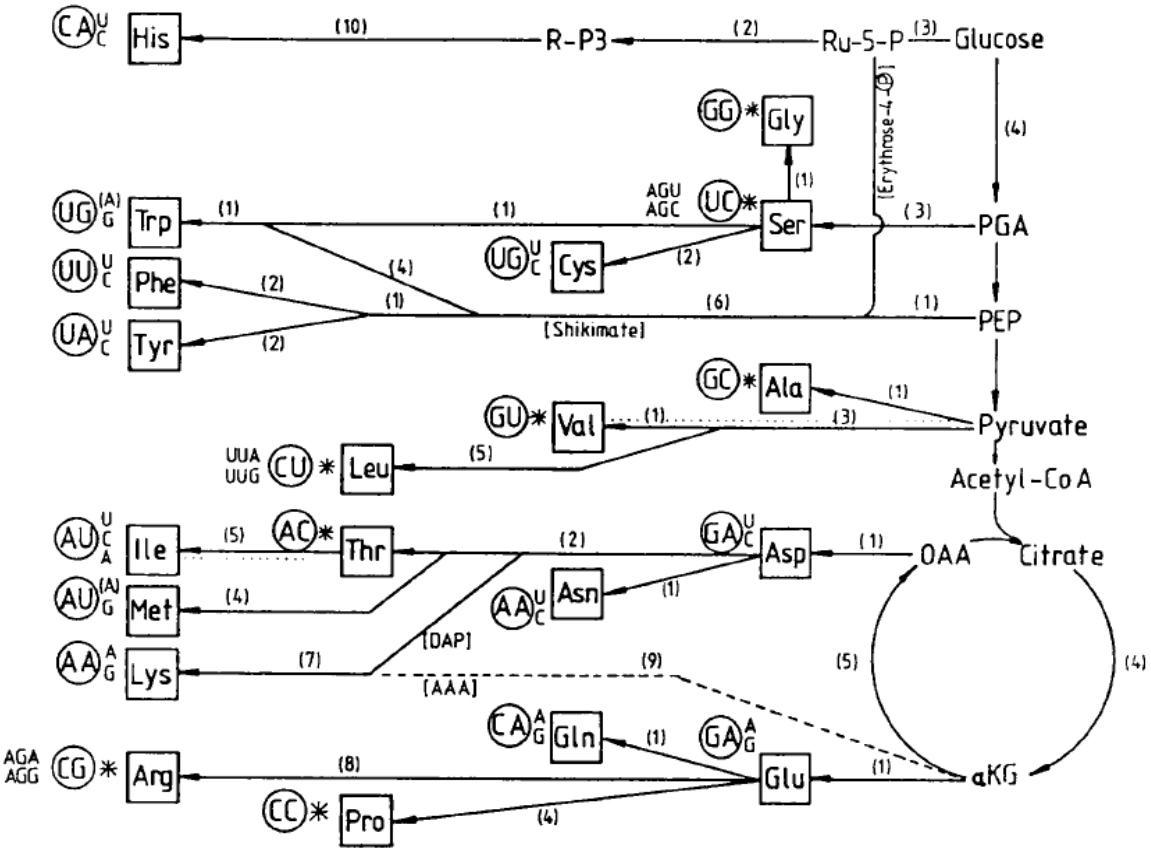

Fig. 3: Rutas de síntesis de los aminoácidos esenciales (Taylor y Coates, 1989)

Otra relación observada es la que existe entre la segunda letra del codón y el aminoácido codificado. Por alguna razón, esta base se encuentra asociada a las propiedades

fisicoquímicas de dicho aminoácido, en particular a su polaridad y en última instancia a su hidrofobicidad o hidrofilia.

Como se puede observar en la tabla 1, la segunda letra de los codones de los aminoácidos más hidrofóbicos es U; en el caso de los aminoácidos más hidrofílicos (es decir, los aminoácidos polares con carga) la segunda letra es A. En el caso de los otros dos nucleótidos (G y C) la relación es un poco más ambigua, pues codifican para aminoácidos con hidrofobicidad intermedia. Existen dos grandes excepciones: la cisteína (hidrofóbica) y la arginina (altamente hidrofílica).

Aminoácido	Índice de Hidrofobicidad	Codones usados
Isoleucina	4.5	A**U***
Valina	4.2	G**U***
Leucina	3.8	C**U*** G**U**R
Fenilalanina	2.8	U**U**Y
Cisteina	2.5	U**G**Y
Metionina	1.9	A**U**G
Alanina	1.8	G**C***
Prolina	1.6	C**C***
Glicina	-0.4	G**G***
Treonina	-0.7	A**C***
Serina	-0.8	U**C*** A**G**Y
Triptofano	-0.9	U**G**G
Tirosina	-1.3	U**A**Y
Histidina	-3.2	C**A**Y
Aspartato	-3.5	G**A**Y
Glutamato	-3.5	G**A**R
Asparagina	-3.5	A**A**R
Glutamina	-3.5	C**A**R
Lisina	-3.9	A**A**Y
Arginina	-4.5	C**G*** A**G**R

Tabla 1: Índice de hidrofilia de los aminoácidos esenciales

Es interesante notar que cuando aunque la guanina tiende a codificar aminoácidos de hidrofobicidad intermedia en la segunda posición, parece ser una especie de comodín. Algunos autores consideran que la guanina, en el "código dentro del código", es un nucleótido informacionalmente neutro. G en la segunda posición no tiene una identidad fisicoquímica definida, pues tiene aminoácidos representativos de todo el espectro de hidrofilia; además cabe mencionar que en los tres aminoácidos con redundancia de 6 codones, dos son consistentes en la segunda letra (Leucina = CUN/GUR; Arginina = CGN/AGR). En el tercero (Serina = UCN/AGY), la pareja "externa" (AGY) tiene G en la segunda posición.

Una última relación entre el código genético y las características fisicoquímicas de los aminoácidos que codifica es, tal vez, la más comprensible. En general, los aminoácidos que tienen un mayor peso molecular tienen una cantidad menor de codones asignados. Esta relación también ha sido mencionada en Taylor y Coates (1989). La razón por la que parece la más comprensible es que un peso molecular menor implica, usualmente, una síntesis más simple; por ello, los aminoácidos de menor peso pueden considerarse más disponible en el momento de una estructuración temprana del código. Incluso, Osawa y Jukes (Osawa et al., 1992) sugieren que los aminoácidos complejos que tienen asignado un único codón fueron (triptofano, metionina) fueron adiciones recientes al código, mediante un mecanismo de "captura de aminoácidos", y solamente después de que se desarrollaron rutas de biosíntesis para ellos.

A simple vista se puede apreciar la misma excepción de los casos anteriores: la arginina, con un peso molecular de 174, tiene asignados seis codones. Es pertinente analizar en este momento las posibles razones (o, al menos, las razones reportadas en literatura) para explicar la consistencia de este aminoácido en presentarse como una excepción a la luz de las regularidades generales del código genético.

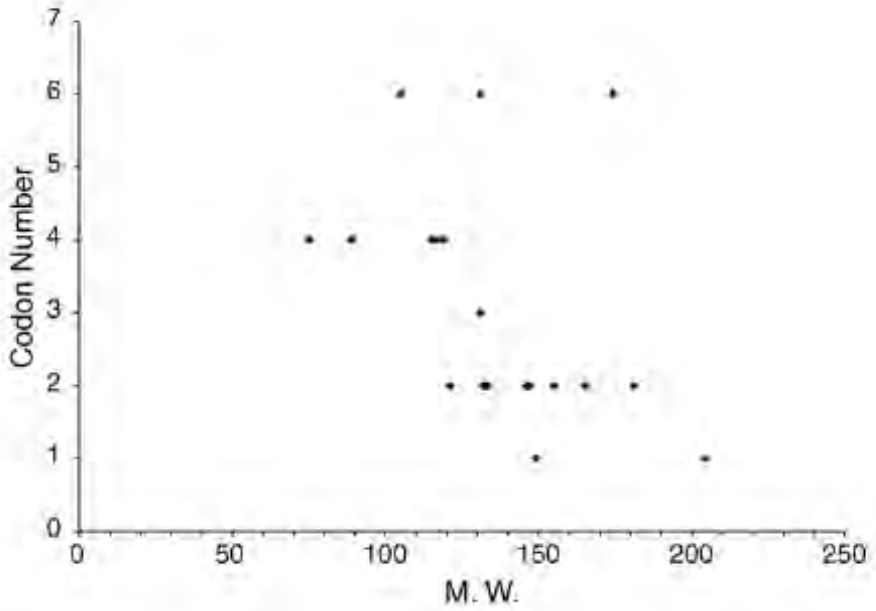

Fig. 4: Correlación entre peso molecular y número de codones. El punto superior izquierdo corresponde a la Arginina. (Di Giulio, 2005)

Los trabajos más relevantes son los llevados a cabo por el grupo que apoya la teoría estereoquímica (Knight 2000, Yarus 2000, 2005). En sus artículos reportan constantemente la selección de ribozimas que se unen a residuos de arginina, *cuyos sitios activos están constituídos precisamente por secuencias de los codones asignados actualmente a la arginina.* Esto apunta a que, al menos en este caso, es posible que exista un factor estereoquímico en la asignación de codones (aunque no existen evidencias similares que relacionen de esta manera a otros codones con sus aminoácidos).

Por otro lado, Di Giulio ha realizado un estudio similar a la comparación de peso molecular con el número de codones, tomando otras propiedades como el índice de termofilia y de barofilia de los aminoácidos (Di Giulio 2005). De esta manera, ha encontrado una correlación directamente proporcional de las propiedades extremófilas de cada aminoácido (comparando los aminoácidos que utilizan preferentemente los organismos extremófilos y aquellos que usan los no extremófilos) y el número de codones que presenta cada uno. La arginina tiene un alto índice de termofilia y barofilia, aunque en los organismos mesófilos dicho aminoácido esté sujeto a fuertes restricciones selectivas. Estos resultados llevan a Di Giulio a proponer que el origen del código (y extrapolándolo

con la propuesta del presente trabajo, el origen de la vida) se llevó a cabo en ambientes de presión y temperatura elevadas.

Todos los factores mencionados anteriormente (hidrofobia e hidrofilia, rutas bioquímicas compartidas, representación diferencial de distintos aminoácidos) hacen que el código genético adquiera propiedades de *resistencia a los errores*. Si un aminoácido de un mensaje es sustituído por otro, ya sea a causa de alguna mutación o por un error de traducción, es altamente probable que el aminoácido nuevo sea similar fisicoquímicamente al original. Esto no se puede considerar como detección o corrección de errores. Sin embargo, otro tipo de regularidades en el código (las informacionales) pueden considerarse como un mecanismo de corrección de errores que no se basa en la similitud fisicoquímica de los aminoácidos cercanos.

Características informacionales.

Así como el código genético presenta patrones estructurales cuando se analizan las características fisicoquímicas de los aminoácidos, un análisis informacional presenta ciertas regularidades que contribuyen a la reducir la carga de errores. Tales regularidades no necesariamente provienen de la identidad fisicoquímica de cada elemento, pero son de importancia fundamental para reducir el impacto de las mutaciones espontáneas que puedan causar daño a los organismos.

La primera de ella, por supuesto, es la llamada "degeneración" del código (a la que en este trabajo se le llamará redundancia). Como se puede comprobar con sólo echando un vistazo a la tabla de codones, muchos de éstos se pueden agrupar en familias por sus dos primeras letras, siendo la tercera indiferente para la traducción. Por ejemplo, todos los codones que comienzan con CC codificarán para el aminoácido Prolina, sin importar la última letra. En otros casos, las agrupaciones se dan por parejas: lo que determina el aminoácido a codificar son las dos primeras letras del codón y la identidad química del último nucleótido (es decir, si éste es una pirimidina o una purina). Por ejemplo: los codones AAR codificarán para Lisina, y los AAY codificarán para Asparagina.

Como ya se ha mencionado, esta característica es la propiedad más evidente que indica un orden en el código genético. Un acomodo aleatorio, incluso si fuera redundante, no produciría estas agrupaciones por familias. Francis Crick propuso poco tiempo después de

la elucidación del código la hipótesis del bamboleo (wobble), la cual se demostró después (Crick, 1966; revisado recientemente en Agris et al, 2007), que sugiere que las primeras dos bases del codón son las más importantes para la codificación, y que el apareamiento de la última base en relativamente débil, por lo que permite una ambigüedad de la relación codón-anticodón.

Existe otra serie de propiedades y patrones relacionados con las estructuras matemáticas del código. La más sencilla es la simetría entre las familias que requieren especificar la última letra del codón y las que tienen ambiguedad completa: cada grupo comprende ocho familias, es decir, exactamente el 50% del código. En una serie de trabajos, Hornos et al (1999, 2004) encuentran otro tipo de simetrías, exactas y globales, presentes en el código. Mediante un mapeo en un octaedro de los elementos del código genético encuentran que existe una simetría de Klein; además, argumentan que esta simetría es congruente no sólo con el código genético estándar, sino con los códigos alternativos encontrados en organelos y procariontes. Los autores sugieren que esta simetría es la causa de la redundancia observada del código, lo cual en nuestro punto de vista posiblemente sea invertir la posible cadena de eventos: el código se debe de estructurar a partir de necesidades informacionales y bioquímicas, y la simetría lograda en último término puede ser sólo un artefacto o uno de varios estados estables posibles. Esta discusión merece un espacio propio, pues no únicamente incluye a las simetrías sino a las regularidades aritméticas (por ejemplo, Shcherbak, 2003), las proporciones áureas (Rakočević, 1998), y otros tipos de patrones; desafortunadamente, cae fuera de los objetivos y temas del presente trabajo.

Las propiedades anteriormente mencionadas, tanto informacionales como fisicoquímicas o matemáticas, tienen un efecto combinado: causan que el código genético tenga propiedades de reducción de errores. Durante los últimos cuarenta años, ha existido una discusión centrada en la posible causa de estos eventos. ¿Es el código genético resistente a errores por mera contingencia, o esta característica ha surgido precisamente gracias a un proceso de selección natural?

En un artículo clásico, Freeland (Freeland y Hurst 1998) investiga la posibilidad de que azarosamente se haya logrado una configuración con tal cantidad de adaptaciones aparentes, y llega a la conclusión de que ésta es mínima. Desde un punto de vista ingenieril,

pueden fabricarse códigos genéticos que sean mucho más eficientes que el utilizado por los seres vivos terrestres actuales en sus propiedades de disminución de errores (Knight 1999; Di Giulio y Medugno 1998, 2001; Di Giulio 2005). Es importante tener en cuenta que no se pretende que la configuración actual es la óptima. Probablemente después de un periodo de establecimiento, las relaciones codón-aminoácido se fijaron, mediante selección, en un pico cercano del paisaje adaptativo. Los resultados de Freeland indican que de un millón de códigos generados aleatoriamente, sólo uno resulta "mejor" que el código natural. Por supuesto, como él mismo lo discute, tales resultados son informativos pero de ninguna manera definitivos: debido a que su técnica asigna pesos a los distintos tipos de mutación, sus parámetros resultan algo aleatorios. Existen muchas maneras de asignar esos pesos; Freeland, por ejemplo, decide ignorar la regularidad fisicoquímica de la segunda base, por lo cual en sus cálculos la eficiencia de cualquier código aleatorio es igual que el código natural, si se considera únicamente la base intermedia. Al considerar la estructura y función de cualquier proteína hipotética, vemos que esto no es así: un cambio de un aminoácido polar por otro polar tendría un efecto menos grave en el plegado protéico o en las regiones transmembranales que cambiar un aminoácido polar por uno alifático no polar. En cualquier caso, el código genético natural resulta significativamente resistente a errores.

Sin embargo, la resistencia a errores no es un argumento definitivo a favor de la construcción del código genético mediante la selección natural. Éste no se consolidó de manera aleatoria, pero las regularidades pueden ser explicadas mediante una serie de restricciones físicas, como se propone más adelante. En efecto, es posible que la cualidad reductora de errores del código sea simplemente un subproducto de tales restricciones. Tan sólo la redundancia de la última letra del codón da cuenta de gran parte de su eficiencia; por ejemplo, en los casos en que la familia de un aminoácido es de cuatro miembros tres mutaciones de nueve posibles resultan silenciosas. En el caso del codón CUA (Lys), cuatro mutaciones de nueve son silenciosas. Si tomamos en cuenta las otras regularidades ya discutidas *y además asumimos que éstas fueron determinantes para la asignación de codones y no surgieron como un mecanismo de optimización a posteriori*, la reducción de la carga de errores se puede considerar como un producto afortunado pero probable, una consecuencia del surgimiento del código por otros mecanismos.

Un escenario hipotético para el origen del código.

Uno de los supuestos centrales del presente trabajo es que el código genético no es un sistema que ha surgido de manera aleatoria. Se puede defender fácilmente este supuesto: el código presenta una serie de regularidades que hacen que el proceso de traducción de una secuencia genética sea resistente a errores mediante la estructura de redundancias de las distintas familias de codones. Además, los aminoácidos codificados por codones cercanos frecuentemente son similares fisicoquímicamente (Taylor y Coates 1989). La estructura particular del código sugiere que un surgimiento azaroso es improbable (Freeland y Hurst, 1998). Por el contrario, las regularidades y características estructurales del código genético hacen pensar que es el producto de un desarrollo gradual hasta su estado actual, en un proceso que lentamente ha optimizado su funcionamiento.

Sin embargo, es insuficiente proponer que dicho proceso de evolución fue dictado únicamente por novedades metabólicas, como el surgimiento de rutas de síntesis de aminoácidos más complejos como el triptofano o la fenilalanina (Wong 1975, 2005; Di Giulio 1999). Es necesario considerar un espectro más amplio de mecanismos, que abarquen desde las posibilidades de un sistema químico hasta las dinámicas sociales que se puede establecer en una población de individuos independientes. Proponemos, a manera de hipótesis de trabajo, un escenario histórico de la evolución del código genético en tres etapas.

1. Etapa fisicoquímica.

El sistema necesariamente inicia en un punto en el que existe una variedad de moléculas utilizables. Antes del surgimiento del código genético, no hay posibilidad de traducción y por lo tanto no aparecen mensajes traducibles. En la primera etapa, las relaciones que puedan surgir se llevan a cabo únicamente por las características moleculares de los elementos. Es posible que dichas relaciones tempranas no sean simbólicas; es decir, que la conexión entre dos mundos, necesaria para el establecimiento de un sistema semiótico (Barbieri, 2003) no se lleva a cabo mediante reglas de convención, sino a través de contacto

físico entre los elementos, sin mediación de adaptadores. Así, las dinámicas tempranas de codificación estarían definidas por las afinidades entre nucleótidos y aminoácidos, o la presencia de mecanismos de síntesis que asocien ambos elementos de la relación simbólica.

Anteriormente (Mercado, 2009) se propuso un mecanismo basado en una serie de reacciones de síntesis de aminoácidos a partir de ácidos tricarboxílicos, propuestas originalmente por Copley et al (2005, 2007). Este sistema de reacciones puede facilitar el establecimiento de un sistema codificante, y su eventual desarrollo puede explicar el surgimiento de adaptadores, pero es insuficiente para explicar muchas de las características del código actual, como la reducción de errores y las simetrías (analizadas anteriormente, en la sección de Regularidades y patrones del código genético). Por lo tanto, es necesario considerar otro tipo de dinámicas que delimiten el establecimiento de la estructura del código genético estándar.

Las relaciones fisicoquímicas entre los elementos no presuponen un sistema codificante. Por lo tanto, es posible afirmar que en la primera etapa no existe un sistema biológico como tal, sino simplemente una serie de reacciones de síntesis y catálisis facilitadas por un sistema químico cada vez más complejo. En parte, este sistema puede corresponder a lo que se conoce como la evolución química prebiótica, pues la constante interacción entre los elementos permite la síntesis de tipos novedosos de moléculas.

2. Etapa informacional.

Las posibles interacciones fisicoquímicas entre los elementos materiales del código genético son fundamentales para determinar la creación de relaciones simbólicas. Sin embargo, existen otras dinámicas que pueden modificar la estructura de los códigos. Como ya se ha señalado, Howard Pattee (2001) afirma que en los códigos naturales coexisten dos tipos de características aparentemente contradictorias: las materiales y las simbólicas. Ambos tipos de características aportan diversos efectos a las dinámicas de codificación, y cada una debe de ser estudiada de manera independiente para entender el código como un sistema completo. Las características materiales responden a las necesidades físicas de los elementos, como las afinidades entre las moléculas de los códigos genéticos y celulares, o las restricciones físicas en sistemas simbólicos macroscópicos. Éste es el tipo de dinámica predominante en la etapa fisicoquímica de establecimiento del código. Por su parte, las

características simbólicas se deben de entender de manera completamente abstracta, pues son derivadas de la arbitrariedad inherente a cualquier sistema simbólico.
Dicha arbitrariedad permite que se pueda fabricar un número enorme de códigos distintos utilizando los mismos elementos. Por ejemplo, el código genético utiliza 64 codones y 20 aminoácidos como conjuntos de significantes y significados, respectivamente. Con esos elementos básicos, la cantidad de códigos posibles es enorme (aproximadamente 1.8446×10^{83}). Sin embargo, sólo un subconjunto reducido es capaz de traducir mensajes conservando características informativas necesarias para que el producto sea funcional.

Como se muestra en la Fig 5., los códigos modifican las características de los mensajes al traducirlos. Por ejemplo, un código completamente redundante (Fig 5a) ocasionaría que la entropía y complejidad de la secuencia se redujeran al mínimo. Por otro lado, un código ambiguo (Fig 5b), cuya traducción varíe probabilísticamente, produciría mensajes impredecibles, lo cual posiblemente tendría un efecto negativo pues los mensajes no serían reproducidos de manera consistente. Por último, un código en el que la aparición de cada aminoácido sea equiprobable (Fig 5c) produciría traducciones con altos niveles de entropía con la mayoría de los mensajes.

Las consideraciones anteriores hacen evidente una particularidad del sistema: los códigos sólo tienen sentido si existen secuencias traducibles, y las secuencias traducibles sólo pueden hacerse funcionales mediante la existencia de un código que las procese. Esta es una reformulación de la paradoja del huevo y la gallina, y, como en ocasiones anteriores, nuestra respuesta es que los códigos y las secuencias deben de haber surgido de manera simultánea. Después de la asociación fisicoquímica descrita en el punto anterior, el primer establecimiento de un código verdadero constituído por relaciones simbólicas hace emerger una serie de características nuevas. Nosotros proponemos que en estas primeras etapas de la estructuración del sistema codificante el factor más importante para la selección de secuencias traducibles y códigos viables sean las propiedades informativas del mensaje y del código. En el caso de las secuencias, estas propiedades son principalmente la complejidad y entropía informacional; en el caso de los códigos, la mayor importancia reside en la redundancia y la corrección de errores.

3. Etapa social.

El código genético posee una serie de características que lo hacen extremadamente robusto. El mecanismo más evidente es el uso de redundancia, que ocasiona que una gran parte de las mutaciones ocurridas al azar sean silenciosas. El efecto de dichas mutaciones sobre la proteína codificada es, en teoría, nulo. También existen otros métodos para minimizar los efectos negativos de la mutación: los aminoácidos de las familias de codones cercanas tienen propiedades fisicoquímicas similares (fig. 6).

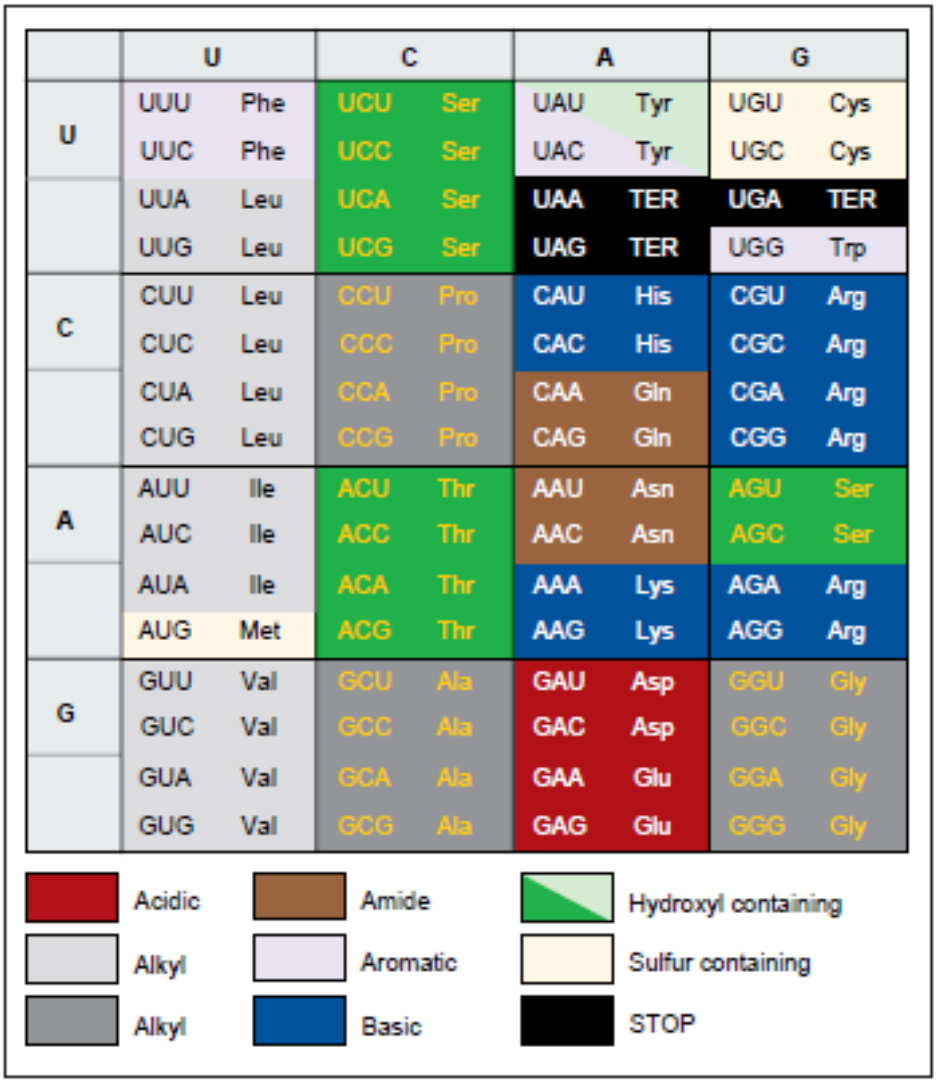

Fig. 5: Patrones fisicoquímicos en la tabla de codones

Esta estructura del código genético debe de haber sido refinada con otros medios, además de los expuestos anteriormente. Ni las afinidades fisicoquímicas ni la conservación

de características informacionales proveen respuestas determinantes acerca de la resistencia a errores. Éste es un fenómeno que implica principalmente una adecuación del organismo como un todo, por lo que se refiere directamente a un proceso de selección y evolución darwiniana.

Sin embargo, ¿es posible considerar otros procesos que complementen a la selección natural para establecer un código genético tan robusto? Es extremadamente difícil reconstruir con precisión el proceso de la evolución de la vida en etapas tan tempranas, pero es factible argumentar algunos principios generales. En el surgimiento de la vida, los mecanismos de control deben de haber sido escasos y laxos, por lo que se puede considerar una gran variacion en diversos procesos del organismo, incluídos los más importantes, como la traducción de la información genética. Por lo tanto, es necesario explorar la posibilidad de que en etapas tempranas de la vida, existía una población individuos, cada uno con un código genético independiente, aunque posiblemente similar a los demás. Además, para que ocurra un surgimiento espontáneo de un sistema complejo codificante tenemos que suponer que tal evento no es completamente improbable: es decir, que en el medio en el que se ha establecido un sistema simbólico, existe una posibilidad no despreciable de aparición de otros, de características similares.

Así, podemos suponer que, en etapas tempranas de la vida, varios sistemas de codificación similares pero independientes coexistían en un mismo medio. A partir de este punto, hay dos caminos posibles. En el primer escenario, cada código genético es completamente independiente a los demás, y cada uno tiene la posibilidad de aumentar su robustez a través de mecanismos fisicoquímicos, informacionales o de selección. Esto significa que el código genético actual es el único que se fijó en las generaciones posteriores a su estructuración, y los demás se extinguieron en las primeras etapas de la vida. En el segundo escenario, los códigos genéticos pueden interpretar mensajes generados en otros contextos; en este caso, los más exitosos son los que tienen la posibilidad de utilizar mensajes funcionales de otros sistemas de manera mínimamente cercana y hacer uso de ellos. En cualquiera de los dos casos, la estructuración y la fijación de un código genético debió de haberse llevado a cabo antes de la diversificación de la vida, i.e., antes del surgimiento de los tres dominios de la vida, puesto que todos los seres vivos del planeta utilizan el mismo código genético (el estándar o universal). Las variaciones excepcionales

que se presentan en algunos procariontes y hongos (Sugita y Nakase, 1999; Suziki et al, 2002) o en organelos como mitocondrias o cloroplastos (Jukes y Osawa, 1990; Kondow et al, 1999; Watanabe, 2010) son novedades evolutivas, ocurridas mucho tiempo después de la estructuración del código estándar. La hipótesis más aceptada es que se derivan de cuellos de botella, en los que el uso de un codón se reduce y el impacto negativo del cambio es mínimo (Osawa et al, 1992). En estas condiciones, pueden ocurrir mutaciones en el tRNA para que acepte otro tipo de aminoácidos, o modificaciones estructurales que cambien las reglas de bamboleo (ver, por ejemplo, el mecanismo propuesto por Suzuki et al, 2011)

La interacción de individuos con diversos códigos y la formación de un lenguaje consenso no es un proceso simple. Para tener una comprensión cabal del fenómeno, es necesario incluir elementos semánticos en el análisis, y los métodos para este tipo de estudios en sistemas naturales son escasos y en ocasiones poco desarrollados. Luc Steels (1998, 2008, entre varios otros textos) ha desarrollado una serie de trabajos para investigar el origen de un lenguaje consenso. Utilizando un conjunto de agentes con lenguajes independientes, muestra que para llevar a cabo la adquisición de símbolos ("symbol grounding", en el término utilizado por Harnad, 1990) es suficiente un significante y una señalización. La señalización es un acto físico, que en el caso de los agentes de Steels consiste en indicar explícitamente al objeto asociado con la palabra. Cuando un agente articula una palabra y otro responde de la manera esperada, se comienza a construir un sistema simbólico compartido.Es evidente que los sistemas simbólicos biológicos no tienen manera de llevar a cabo ese tipo de enseñanza y aprendizaje, pero el proceso de señalización se puede generalizar. En este sentido, una traducción de una molécula funcional es equivalente a una señalización correcta. Así, dos códigos compatibles podrán hacer uso de una serie de mensajes compartidos. Al aumentar el número de códigos y de mensajes, las restricciones de compatibilidad serán mayores.

Las tres etapas pueden considerarse parte de una secuencia cronológica. La primera no presupone más que una serie de moléculas con características fisicoquímicas diversas, lo cual se puede encontrar en casi cualquier sistema químico. La segunda presupone un mensaje simple, con posibilidades de evolución abierta y de coevolución con el código que lo interpreta. La tercera implica una población de individuos independientes, con códigos

desarrollados, que poseen la capacidad de comunicarse entre sí y construir convenciones sociales. El presente trabajo se enfoca en la etapa de estructuración informacional. Los parámetros que se explorarán se refieren casi exclusivamente a la capacidad de transmitir un mensaje, interpretarlo, y conservar características propias a las secuencias de símbolos, como complejidad algorítmica o entropía informacional. Por otro lado, algunas interrogantes importantes quedan fuera de los límites de esta investigación, como el origen de los primeros mensajes codificantes, o la evolución del repertorio específico de elementos del código, tanto nucleótidos como aminoácidos.

4. El experimento

Después de exponer todas las características del tipo de investigación que se quiere realizar, la justificación de ésta y el marco hipotético dentro del que se mueve, presentamos un experimento informcional. En él, intentamos explorar las capacidades de procesamiento de información del código genético actual, y compararlas con otros códigos, generados de manera aleatoria. El experimento toma un camino similar al artículo clásico de Freeland y Hurst (1998), en el que analizaban la optimización del código genético de una manera ingenieril, enfocándose en la reducción de errores; sin embargo, nuestro acercamiento difiere radicalmente, pues no nos interesan las posibles características de optimización. De hecho, los códigos individuales son mucho menos importantes que el procesamiento que hagan de la información. El terreno que exploramos se define por la comparación de las características informacionales de una cadena constante y aquellas de las cadenas resultantes, después de llevar a cabo el proceso de traducción. Los códigos que arrojan resultados de interés no necesariamente serán más eficientes; únicamente procesarán la información biológica de manera similar a los sistemas vivos. Al no tomar en cuenta otros factores, como la naturaleza fisicoquímica de los elementos del código genético, la reducción de errores o la información semántica que existe en los genomas, podemos analizar directamente la estructura de codificación que permite conservar un determinado nivel de complejidad informacional.

Método

Para la generación y análisis de los códigos genéticos alternativos de manera eficiente, utilizamos una serie de plataformas de programación y software. La producción de los datos puede dividirse en varias etapas independientes:

Generación de códigos:

Los códigos genéticos alternativos se generaron a través de un modelo programado en Netlogo, una plataforma con interfaz gráfica basada en agentes.

1. El modelo está diseñado para admitir cierto número de variantes. Puede ser configurado para que genere códigos con diverso número de codones de término a los que no se les asocia ningún aminoácido, y es posible cambiar el número de codones, aminoácidos y códigos generados.
2. En el primer paso (*Setup*), Netlogo crea dos poblaciones de agentes. Una de ellas representa el conjunto de codones del código genético; es decir, las 64 combinaciones que se pueden llevar a cabo con los cuatro nucleótidos del DNA o RNA. La otra representa a los 20 aminoácidos esenciales del código genético estándar. Debido a los objetivos del proyecto, los números de aminoácidos y codones se mantuvieron fijos; además, a cada agente se le asignó un codón o aminoácido específico, y un valor aleatorio de "afinidad" entre 0 y 1 como variables propias.
3. En el segundo paso (*Go*), los agentes se mueven de manera aleatoria en el espacio de interacción. Cada vez que un aminoácido y un codón se encuentran, comparan sus afinidades, y la probabilidad de asociación se calcula por la siguiente fórmula:

$$1 - |A_A - A_c| = p(A),$$

en donde A_A es el valor aleatorio asignado al aminoácido en cuestión, A_c es el valor aleatorio del codón, y p(A) es la probabilidad de asociación entre ellos. Los dos valores de afinidad se comparan y la diferencia entre ellos (es decir, el valor absoluto de su resta) determina qué tan probable la creación de un enlace entre ellos. Esto quiere decir que entre más cercanos sean los valores de A_c y de A_A, más alta será la probabilidad de asociación simbólica entre ellos. Finalmente, se utiliza un valor generado aleatoriamente para simular un proceso probabilístico y decidir si en efecto se asociarán o no. Los codones que ya han sido relacionados con un aminoácido no se pueden relacionar con ningún otro, para evitar un código ambiguo; sin embargo, la redundancia es posible e incluso inevitable.
4. Al acabar de asignar los 64 codones a sus respectivos aminoácidos, el modelo produce dos archivos de salida: uno, escrito en el formato de un diccionario de Python, enlista todo el código; el otro especifica la redundancia de cada aminoácido, es decir, el número de codones que codifican para cada uno.

Traducción de secuencias.

Los códigos generados en Netlogo se utilizaron, junto con el código standard, como diccionarios para la traducción de secuencias genómicas reales de organismos de distintos dominios mediante una aplicación programada en el lenguaje Python 2.6. Dicho programa acepta archivos de texto con secuencias de DNA y los traduce con el código genético que se le indique, produciendo una secuencia de aminoácidos en un archivo de salida.
Para el análisis de los códigos se tradujeron genomas completos, pues el análisis de características informacionales de una secuencia se hace más preciso entre más grande sea la muestra analizada. A continuación se detallan los datos de los genomas utilizados.

Organismo	Dominio	Bases en genoma	Número de bases codificantes	Porcentaje de información codificante	Número de genes (genes codificantes)
Mimivirus de *A. polyphaga*	Virus (dsDNA)	**1,181,549**	1,039,763	88%	1018 (979)
Mycoplasma genitalium	Bacteria	**580,076**	522,068	90%	524 (475)
Nanoarchaeum equitans	Archaea	**490,885**	446,705	91%	586 (540)
Schizosaccharomyces pombe	Eukaryota	**2,452,883**	1,226,441	50%	1081 (898)

Tabla 2: Características de los genomas utilizados en el experimento

Análisis de contenido informacional.

Para comparar las secuencias producidas al traducir un genoma por medio de los diversos códigos, era necesario establecer una medida informacional cuantitativa cuyo valor fuera rápidamente comparable con los de otras secuencias y que reflejara la estructura de la secuencia de aminoácidos. Optamos por utilizar un método estadístico de compresión publicado por Cao et al. (2007). Más adelante, en la primera sección de la discusión, se discutirá cuáles son las principales diferencias entre los parámetros informacionales que se pudieron utilizar, y por qué el método de compresión es el más cercano a las necesidades del proyecto. También se describirá someramente el algoritmo de compresión desarrollado por Cao et al.

Utilizando este algoritmo, se comprimieron los genomas completos y las traducciones realizadas por cada código generado por el programa de Netlogo. Asimismo, se realizaron traducciones con el código genético estándar y con un código para producir una secuencia completamente repetitiva, es decir, con un código de redundancia máxima en el que todos los codones se asocian con el mismo aminoácido. Esta secuencia funciona como control, pues su compresibilidad es máxima y, por lo tanto, su complejidad es mínima.

Resultados

En el experimento, los resultados producidos por el software eXpertModel toman la forma de valores positivos, cuyas unidades son bits por símbolo (bps). El intervalo de valores posibles varía dependiendo del tamaño del alfabeto utilizado en la secuencia, y es posible obtener el valor máximo con la fórmula

$$\log 2\ A = vmax,$$

en donde A es el número de símbolos que componen el alfabeto. En el caso de las secuencias de ácidos nucléicos, el alfabeto se compone por cuatro símbolos: [A, G, T, C], en el caso del DNA, y [A, G, U, C] en el caso del RNA. El valor máximo de información es de 2 bps. Por su parte, las secuencias protéicas se componen de 20 aminoácidos, y el valor máximo de información es de 4.322 bps (aprox).

Así, para cada genoma obtuvimos 1000 valores distintos de contenido informacional, uno por cada traducción. La distribución de frecuencias de dichos valores se muestra a continuación.

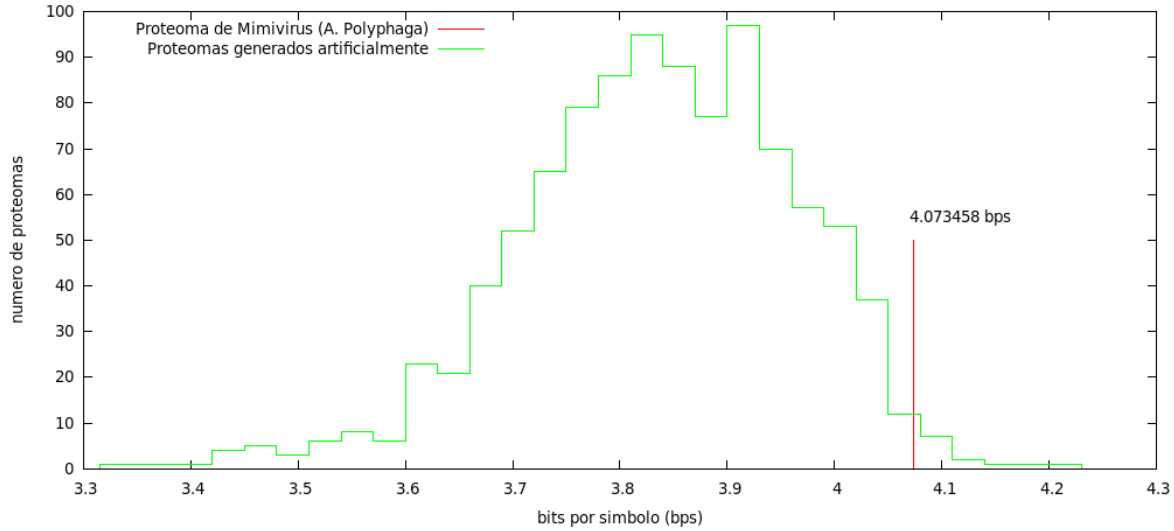

Fig. 6: Distribución del contenido informacional de los proteomas generados con el genoma de Mimivirus

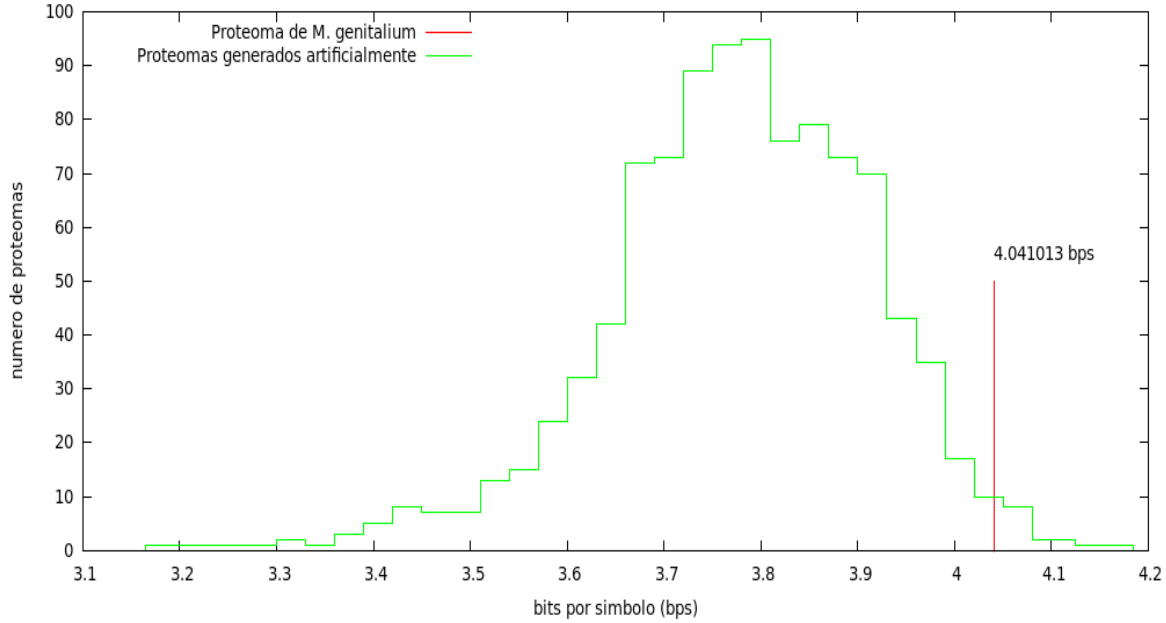

Fig. 7: Distribución del contenido informacional de los proteomas generados con el genoma de M. genitalium

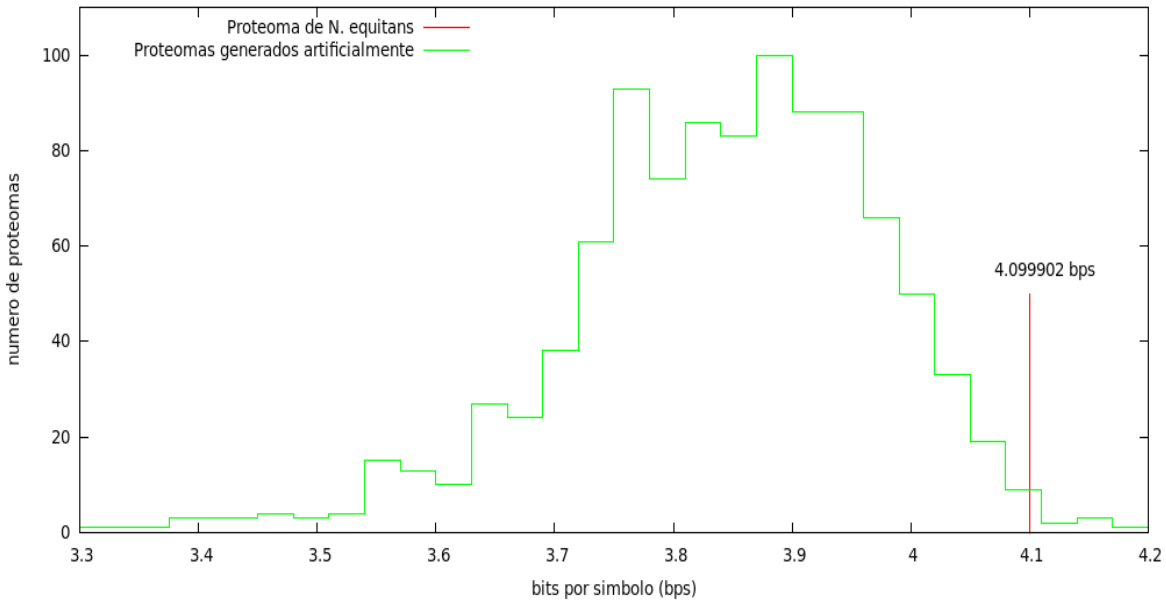

Fig. 8: Distribución del contenido informacional de los proteomas generados con el genoma de N. equitans

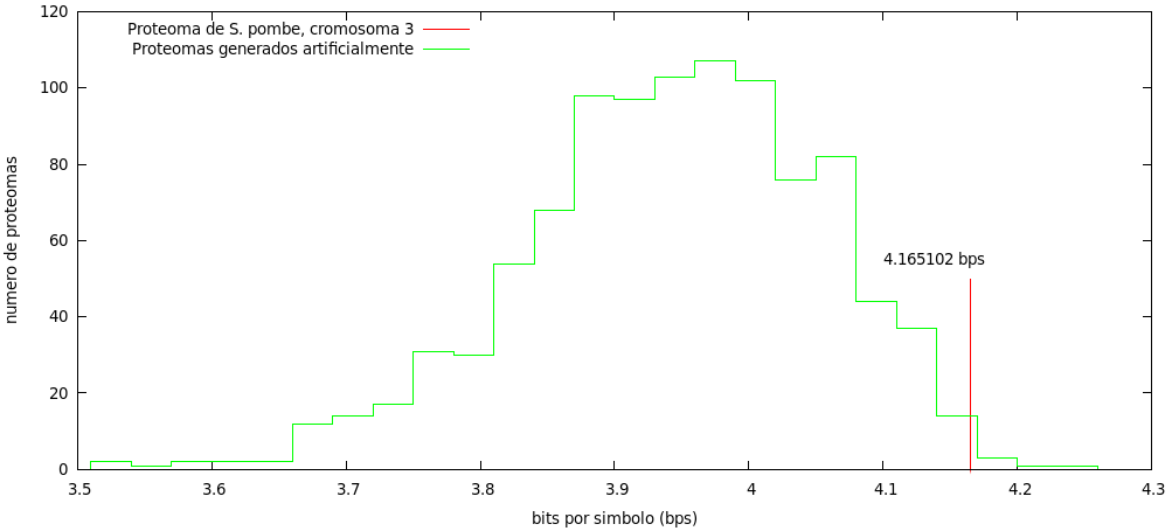

Fig. 9: Distribución del contenido informacional de los proteomas generados con el genoma de S. pombe

En las gráficas anteriores se muestran los valores informacionales de los proteomas obtenidos utilizando el código genético estándar con una línea roja (con la leyenda "Proteoma de *M. genitalium*", " de *N. equitans*", "de *S. pombe*" o "de *Mimivirus*"); dichos valores sirven de referencia para comparar las compresiones de todas las demás traducciones, las cuales se muestran en verde bajo la leyenda de "Proteomas generados artificialmente". El eje x representa el grado de compresión del proteoma obtenido: cerca del origen se encuentran los proteomas con menos bps (y por lo tanto más compresibles); al lado izquierdo de la gráfica, se encuentran los valores cercanos al máximo informacional (4.332 bps), y por lo tanto los proteomas encontrados en esta región sólo pudieron ser comprimidos en un grado muy pequeño. Como se puede observar, el contenido informacional de los proteomas generados artificialmente se distribuyen de manera normal alrededor de un promedio que se aleja significativamente del valor del proteoma "original", obtenido con el código genético estándar. Un análisis estadístico un poco más profundo se realiza en la sección "Acerca de los resultados", de la parte 5: Discusiones.

En la sección de Discusiones también se discuten las implicaciones biológicas de estos resultados. Las gráficas contienen, de manera implícita, una serie de datos acerca de la naturaleza del procesamiento biológico de información. Si bien en ellas se condensan datos generados artificialmente con mensajes encontrados en un sistema celular real, ambos son comparables pues han sido traducidos por sistemas de codificación análogos. Esta comparación da pie a una discusión extensa, tanto física como biológica e incluso metodológica, que se detalla a continuación.

5. Discusiones

La primera parte de esta tesis es una extensa discusión introductoria antes de exponer el experimento realizado. En ella, tratamos diversos temas que creímos necesario entender antes de tratar el experimento informático. Por ejemplo, se hace un recuento del estado del campo de la biosemiótica, que es el soporte central de varios de los argumentos, intereses e incluso supuestos que dieron forma al trabajo; también se explica el escenario hipotético del origen del código genético, del cual se toma sólo una parte para la presente investigación, y los conceptos generales que se utilizan para el análisis. Después de presentar los resultados experimentales, es necesario otro tipo de discusión. El primer punto que se debe discutir es cómo se eligió la metodología y los parámetros que hemos utilizado. Creo necesario discutir también qué otras opciones existen para la medición de la información biológica, y cuáles son las diferencias y similitudes tienen con nuestro método.

Acerca del método.

Tal vez resulte extraño realizar un trabajo acerca de la información de una secuencia, y no utilizar las medidas derivadas de la teoría de la información. Hemos afirmado anteriormente que una deficiencia (o fortaleza, dependiendo del punto de vista que se tome) de la teoría de Shannon es el rechazo de todo contenido semántico del mensaje, enfocándose a la comunicación literal. La presente investigación parece tomar el mismo rumbo: no hemos tratado de discernir las funciones de cada una de las partes del mensaje de cada uno de los genomas, y, en efecto, ni siquiera hemos tratado de ver si existen funciones similares entre ellas. Dichas investigaciones quedan fuera de los objetivos de este trabajo, y no se han tomado en cuenta. Entonces, ¿por qué no utilizar la entropía informacional, la cual es uno de los conceptos más eficientes y refinados para medir comunicación literal, para realizar las mediciones de interés?

Nosotros tuvimos que salir del campo de la teoría de la información por un par de razones que hacen que la entropía informacional sea inconveniente para nuestro estudio. Si bien es

cierto que no tomamos en cuenta la función de cada mensaje analizado, una de nuestras premisas principales es que los mensajes poseen contenido semántico. Esto quiere decir que no nos importa qué funciones pueda llevar a cabo cada mensaje, pero es fundamental poder suponer que tal contenido existe e incluir la presencia o la ausencia de semántica en nuestro análisis. A partir de ello, realizamos la hipótesis de que la semántica, en general, y la información biológica con contenido funcional, en particular, tienen características informacionales determinadas. El argumento principal de este trabajo también se deriva de dicha argumentación: *suponemos que los códigos biológicos y las cadenas que éstos traducen se afectan mutuamente, y gran parte de las características de ambos está determinada por esta relación bidireccional.* Es decir, un mensaje funcional de DNA sólo puede ser traducido a un mensaje funcional en el lenguaje de las proteínas por un código que tenga cierta estructura, en particular en términos de redundancia; por su parte, una vez establecido el código, los mensajes funcionales de DNA deben estar acoplados a él.
La teoría de Shannon no permite incluir estas consideraciones en un análisis. Uno de los supuestos fundamentales es que todas las cadenas de símbolos son equivalentes. Cada una representa un solo punto en el espacio de posibilidades de las cadenas de determinada longitud, formadas con un alfabeto de determinado número de símbolos. Como ya se ha afirmado muchas veces antes, la semántica, y por extensión la funcionalidad, no tienen cabida alguna en el sistema de la teoría ingenieril de la información. Es cierto que es posible discernir entre dos cadenas distintas de caracteres: una cadena muy repetitiva tiene una entropía informacional baja, mientras que el valor de esta medida es cercana al máximo en cadenas de cDNA, pues cada uno de los nucleótidos aparece más o menos el mismo número de veces (la proporción exacta, por supuesto, varía entre distintos organismos). Sin embargo, en el continuo de valores de entropía es imposible incluir alguna referencia que sirva para discernir entre *tipos* de información. Existen dos artefactos teóricos que demuestran este impedimento, ambos relacionados directamente con el rechazo de la semántica.

En primer lugar, la teoría de la información es ciega al *orden* de las letras que aparecen en una secuencia. Debido a que las únicas cantidades utilizadas en la fórmula de la entropía informacional son el número de veces que cada letra aparece en la secuencia, y la longitud de ésta, no hay manera de incluir datos acerca del orden de aparición de cada

caracter. Por tal razón, una cadena determinada y cualquiera de las permutaciones posibles de sus letras tienen exactamente la misma entropía. Como ejemplo, considérese la cadena de letras ABCABCABCABC. Ésta tiene la misma entropía que las cadenas AAAABBBBCCCC y AABACBBACCCB, aunque la primera de las tres tenga un patrón repetitivo tres letras (ABC), la segunda enuncie cada letra cuatro veces y la tercera no tenga un patrón reconocible a primera vista. Así, la entropía informacional no sólo ignora la semántica, sino que descarta todo análisis de la estructura del mensaje. Esto puede ser parcialmente resuelto utilizando entropía condicional. Por ejemplo, Hubert Yockey (2005) utiliza entropía mutua para calcular la similitud entre diversas proteínas homólogas, lo cual a su vez usa para determinar aminoácidos que son funcionalmente equivalentes. Claramente, sus investigaciones representan una manera de hacer que un valor no semántico (la entropía informacional) nos provea de datos completamente semánticos (la equivalencia funcional entre aminoácidos de proteínas homólogas). No obstante, existe otro problema que hace que la teoría de Shannon sea inadecuada para este estudio.

El segundo problema es la definición de lo que se va a utilizar como símbolo al realizar el análisis. Es necesario realizar dicha definición *a priori*, pues distintos tipos de símbolo arrojan resultados muy distintos. Una de las implicaciones más profundas de este hecho es que un símbolo no necesariamente tiene que ser un carácter. Por ejemplo: las cuatro letras de una cadena de RNA pueden expresarse con cuatro caracteres distintos (i.e., A, G, C y U). Sin embargo, dichas letras también pueden ser traducidas a un sistema binario, en el que cada una necesita de dos dígitos para ser expresada; es decir, los cuatro símbolos de ese alfabeto son “00”, “01”, “10” y “11”. Así, en tal caso un símbolo independiente consta de dos dígitos binarios y un único dígito aislado no tiene ningún sentido ni acarrea consigo información. Para entender la importancia de este hecho en nuestro análisis, considérese la cadena de letras del párrafo anterior: ABCABCABCABC. Si definimos a los símbolos como letras aisladas, el mensaje sería A-B-C-A-B-C-A-B-C-A-B-C. Este alfabeto consta de tres símbolos y su entropía informacional (calculada con la sumatoria de $p(x)\log_2 p(x)$) es de aproximadamente 1.585. Sin embargo, si se define al símbolo como una pareja de letras, el mensaje sería AB-CA-BC-AB-CA-BC. Como se puede observar, el mensaje se vuelve mucho más corto, y aunque la entropía informacional se conserva en el mismo valor, no se utilizan todas las combinaciones posibles de letras: no

están presentes los símbolos AC, CB ni BA. Si, en cambio, se utilizan símbolos de tres letras, el mensaje sería ABC-ABC-ABC-ABC. En este caso, la cadena consta de un solo símbolo por lo que la entropía informacional es cero (pues se considera que log2(1) = 0). Como ya se ha dicho, en el campo de la ingeniería esta rigidez resulta ventajosa: se evitan consideraciones de ambigüedad y de interpretación subjetiva, se tiene un marco sólido para analizar un mensaje y los resultados que arroja un análisis de Shannon respecto a una determinada cadena de letras son interpretables de manera unívoca y casi inmediata. Nosotros debimos escoger otra medida más amplia (y necesariamente más complicada) para poder incluir diversos tipos de patrones y efectos del contenido semántico en los mensajes y los códigos.

Otra medida relacionada con la información de un mensaje es la complejidad de Kolmogorov. Con ella, Kolmogorov (1965) y Chaitin (1969), quienes la propusieron por primera vez de manera independiente, tratan de acercarse a los aspectos que deja fuera la teoría de Shannon. El concepto detrás de ella es relativamente simple: la complejidad de una secuencia es equivalente a la longitud mínima de su descripción. Por ejemplo, una cadena de cien "A" repetidas tendrá una complejidad de Kolmogorov reducida. Una descripción probablemente mínima de ella es "A*100"; con esta expresión, una cadena de 100 letras queda reducida por una equivalente de 5 caracteres. Por otro lado, una cadena generada aleatoriamente necesitará enunciarse letra por letra, pues es altamente probable que no contenga patrones reconocibles, y cualquier descripción alternativa posiblemente sea más larga que la cadena misma. De hecho, uno de los intereses originales de la complejidad de Kolmogorov era proveer un concepto congruente de la aleatoriedad. Probablemente se haya notado que las descripciones reducidas de la cadena están basadas en procedimientos algorítmicos; es decir, instrucciones que indican paso a paso cómo obtener la cadena de interés. Considérese el ejemplo del párrafo anterior, donde el algoritmo dirá "Imprimir la letra 'A' cien veces". Es por eso que la complejidad de Kolmogorov es una medida de complejidad algorítmica, y también es por eso que a veces el concepto se describe informalmente como "el programa de computadora más corto que genere como *output* la cadena de interés". Aquí surgen los problemas del uso de ese concepto, que son justamente aquellos que las medidas de Shannon evitan al ignorar a la semántica.

Por un lado, al necesitar explicar los pasos del algoritmo, es necesario utilizar un lenguaje determinado. Esto parece trivial, pero el simple hecho de expresar algo en un lenguaje abre la puerta a dificultades variadas e inevitables, así como preguntas de resolución complicada. Por ejemplo: ¿qué tipo de lenguaje se va a utilizar? ¿Qué tanta libertad tenemos de inventar términos y redefinir palabras? ¿Qué ocurre cuando hay oraciones que sean autorreferentes y se muevan en un plano metalingüístico? En estos problemas se comienza a vislumbrar la presencia de los teoremas de Kurt Gödel (ver Gödel, 1992, una traducción inglesa del artículo clásico en alemán de 1932), que demostró que cualquier sistema lingüístico es incompleto o inconsistente; la repercusiones de ello llegan incluso a las preocupaciones de Ludwig Wittgenstein (1988) cuando señala la dificultad (o, de hecho imposibilidad) de definir inequívocamente incluso los términos más sencillos. Utilizar la complejidad de Kolmogorov implicaría construir, o al menos definir, la estructura y las reglas básicas del lenguaje que utilizaremos para describir la secuencia de letras que nos interesa analizar.

Por otro lado, y a diferencia de la entropía de Shannon, la complejidad de Kolmogorov no es un valor fijo ni fácilmente obtenible. En parte por las dificultades lingüísticas señaladas en el párrafo anterior, y en parte porque no hay un método único de obtener el "programa de computadora más corto" que produzca la secuencia analizada, lo más que se puede obtener en el caso de la complejidad de Kolmogorov es un límite superior que está siempre sujeto a cambios. El límite superior inicial es, justamente, el algoritmo que enuncie letra por letra la cadena de símbolos. En caso de encontrar una descripción más corta, ésta funcionará como límite superior, hasta que se encuentre una descripción aún menor. Incluso en el caso de las descripciones más triviales (por ejemplo, una cadena consistente en la repetición de una sola letra) se presenta una serie de consideraciones que damos por sentadas, como la base numérica que se utilizará para enunciar las cantidades, o las convenciones simbólicas de operaciones aritméticas.

Frente a estas dificultades metodológicas, tuvimos que buscar e implementar un proceso que nos permitiera analizar las secuencias biológicas, tomando en cuenta todos los requisitos anteriores. El método debía de arrojar un resultado simple en forma de un valor numérico, que se permitiera comparar fácilmente las características de diversas cadenas. Además, tenía que ser un proceso repetible y que fuera consistente, sin importar cuántas

veces se volviera a realizar la prueba; tenía que poder aplicarse a cualquier cadena que se tuviera que analizar, independientemente de su estructura. Asimismo, la técnica debía reconocer patrones de letras individuales, como el conteo total de una letra determinada en la cadena, y también patrones locales, como grupos de letras reconocibles que se repitan en diversos lugares del mensaje. Además, un método ideal debía incluir de cierta manera los posibles efectos de un significado funcional de un mensaje, sin caer en las ambigüedades hermenéuticas que evita la teoría de Shannon.

La compresión

Nos parece que la compresión, el método elegido para realizar el análisis, se ajusta a las necesidades expuestas en el párrafo anterior. La compresión de secuencias biológicas es posible porque cualquier cadena es representable como una concatenación de símbolos, utilizando un alfabeto abstracto de longitud *l*. Como se puede ver a simple vista, el principio conceptual detrás de nuestra elección es más cercano a la complejdad de Kolmogorov que a la entropía informacional: las cadenas altamente repetitivas tendrán una tasa alta de compresión, pues grandes secciones del texto se podrán representar con pocos símbolos; las cadenas aleatorias no se podrán comprimir

Así, es posible procesar las cadenas biológicas como archivos de texto electrónico, con los diversos métodos y programas de compresión de archivos electrónicos que existen, desde compresores genéricos como WinZip y gzip hasta aplicaciones con objetivos específicos. Cada uno de estos métodos cuenta con un principio operacional distinto, y tiene diversas ventajas y desventajas. Aunqueéstas no serán discutidas en este texto, conviene señalar que las particularidades de los diversos procedimientos de compresión los hace adecuados para un tipo distinto de información. Éste es el punto en el que la técnica utilizada en este trabajo difiere de la complejidad de Kolmogorov, que sólo nos puede proveer con un valor máximo absoluto. El método de compresión se debe elegir en función al tipo de información que se va a manejar. Distintos algoritmos pueden dar resultados muy variables, y si el método no es el adecuado, la cadena “comprimida” puede resultar incluso más larga que la cadena original.

Esta particularidad metodológica está justamente en la base de la controversia reciente acerca de la posibilidad de comprimir secuencias biológicas. Nevill-Manning y

Witten (1999) afirmaron terminantemente, desde el título de su publicación, que las proteínas no son compresibles. Su análisis abarca dos métodos de compresión. Por un lado, comentan los que construyen diccionarios, como el método de Lempel-Ziv, y los descartan rápidamente debido a la poca repetición de las secuencias protéicas. Por el otro, implementan un método propio, basado en predicciones estadísticas de procesos de Markov. Como una de sus premisas es la posibilidad de encontrar cualquier tipo de mutaciones en las proteínas, no sólo utilizan la información de la cadena que están analizando, sino que dan pesos a las probabilidades de predicción mediante información contextual de datos físicos. Así, introducen en su modelo datos de cercanía de los codones de aminoácidos y de propiedades fisicoquímicas (por ejemplo, los grupos aromáticos de la fenilalanina, la tirosina y el triptófano, el tamaño extremadamente reducido de la glicina, los átomos de azufre de la cisteína). El resultado, contrario a lo que cabría esperar inicialmente, no presenta ninguna mejoría en comparación con los compresores genéricos; de hecho, algunas de las cadenas tienen más bits por símbolo después de la compresión. Por lo tanto, Nevill-Manning y Witten llegan a la conclusón que da nombre a su artículo: las proteínas no se pueden comprimir.[5]

Después de la publicación del artículo de Nevill-Manning y Witten se publicaron diversos trabajos demostrando la posibilidad de comprimir las secuencias biológicas. Hategan y Tabus (2004) reportaron su compresor (ProtComp) en un artículo cuyo título es una referencia directa a "Protein is incompressible": "Protein is compressible". Este método se basa en compresión mediante la elaboración de un diccionario, de manera un poco similar al algoritmo de Lempel-Ziv. La diferencia fundamental es que ProtComp lee, en un primer paso, la cadena a comprimir, e identifica las secciones parcialmente repetitivas; por su parte; Lempel-Ziv construye el diccionario de compresión desde que comienza a leer la secuencia por primera vez. El otro tipo de compresores utiliza predicciones estadísticas para construír la distribución de probabilidades de aparición de los símbolos en la secuencia; algunos de éstos son los algoritmos ARM (Allison et al, 2000; Stern et al 2001) y CDNA (Loewernstern y Yianilos, 1999). Todos han logrado resultados mucho más eficientes que los resultados pesimistas de Nevill-Manning.

5 De hecho, los autores proponen que ninguna cadena biológica funcional se puede comprimir, pero el DNA puede resultar compresible en organismos que tienen una cantidad elevada de DNA no codificante, el cual muchas veces tiene fragmentos repetitivos; debido a ello, el DNA es compresible *ocasionalmente*.

El algoritmo utilizado por nosotros fue publicado por Cao et al (2007). Se trata de un algoritmo completamente estadístico, que agrupa un panel de "expertos", es decir, "cualquier cosa que genere una distribución de probabilidades razonablemente buena para cada una de las posiciones de la secuencia". Con dichas distribuciones, cada uno de los expertos hace una predicción para cada uno de los lugares de la secuencia, y la distribución final que servirá para comprimir la secuencia será la combinación de todas las predicciones. El algoritmo es más complejo de lo que se esboza aquí. Por ejemplo, los expertos tienen distintos valores dependiendo de la corrección de sus predicciones, y a diferencia de la mayoría de los compresores puramente estadísticos, éste utiliza información contextual para realizar las predicciones: la distribución de probabilidades se construye con los 512 símbolos anteriores, lo cual le da más poder predictivo al algoritmo pues una secuencia de DNA puede ser informacionalmente muy heterogénea.

La compresión que hemos realizado en este trabajo no tiene como fin un almacenaje o procesamiento más eficiente de archivos electrónicos. Fue utilizada como una medida de información y complejidad de nuestras secuencias. No es intercambiable con ninguno de los parámetros ya discutidos, aunque se acerca a ambos en algunos puntos. Conceptualmente, se parece más a la complejidad de Kolmogorov, pues se basa en la expresión de una secuencia con una secuencia equivalente de menor longitud, y esta última será una medida cuantificable del contenido informacional de la cadena original. Metodológicamente, es similar a la entropía informacional pues produce un resultado numérico determinado para cuantificar dicho contenido, e incluso las unidades son las mismas (bits por símbolo), y siempre se realiza con un mismo algoritmo (aunque no sea el mismo que el propuesto por Shannon). Sin embargo, a diferencia de la entropía informacional, nuestro método toma en cuenta la estructura de la información, la cual está determinada por el posible contenido semántico de ésta. Un mensaje con ciertos módulos o palabras[6] repetitivas se podrá comprimir más que uno en el que la misma cantidad y tipo de letras aparezcan de manera impredecible, aunque tengan la misma entropía informacional. Por ello, la entropía de Shannon y los resultados de nuestro análisis son comparables mas

6 El término "palabras" se debe de entender como lo que es: una metáfora. De hecho, la interpretación literal de este término al ser aplicado en cadenas de DNA es uno de los errores más frecuentes en el análisis informacional. Al comparar la información biológica con un lenguaje humano (por ejemplo, el idioma inglés) puede pensarse que la primera no tiene ningún tipo de estructura. A estas alturas es evidente que la tiene, pero no es parecida a la estructura de un texto escrito en inglés: es completamente idiosincrática y tiene que analizarse con sus propias reglas y particularidades.

no equivalentes; esta relación se debe de tener en cuenta en la discusión de los resultados, que se presenta a continuación.

Acerca de los resultados.

Aunque hemos utilizado la misma medida informacional, los resultados que hemos obtenido a lo largo de este trabajo se derivan de objetos muy distintos. Por ejemplo, se han analizado las cadenas de cDNA, que se supone que contienen únicamente información con contenido semántico. También existe una gran diferencia entre la información expresada en las cuatro letras de las secuencias de los ácidos nucléicos, y la de las proteínas, que tienen 20 aminoácidos posibles. Asimismo, es necesario tomar en cuenta que estamos analizando tanto información "real", es decir, cadenas que tienen un significado biológico y que son producidas en la célula para llevar a cabo una función específica, como cadenas que hemos producido nosotros con códigos generados computacionalmente, y que aunque tienen las mismas características informacionales, no tienen ninguna función. A pesar de esta serie de disparidades entre los distintos resultados, es posible articularlos enfocándose en los puntos que comparten, tanto metodológica como conceptualmente.

El primer resultado a discutir es la relación entre los resultados de la compresión del cDNA de los organismos y los de la compresión del proteoma completo, traducido utilizado el código genético estándar. Algunas de las características de cada cadena difieren entre una y otra; por ejemplo, el número de letras en el alfabeto de cada tipo, lo cual a su vez afecta el valor máximo de complejidad de la cadena (2.0 para el DNA y el RNA, y 4.322 para las proteínas); también es distinta la longitud, pues las proteínas son tres veces más cortas que las cadenas de ácidos nucléicos. Sin embargo, ambas medidas son totalmente comparables, pues fueron obtenidas utilizando el mismo algoritmo, y las unidades en que los valores están expresados son las mismas. La comparación de los resultados se muestra en la tabla 3. Todos los valores numéricos se encuentran en bits por símbolo.

Organismo	Compresión de DNA	Compresión de DNA (Normalizado)	Compresión de Proteína	Compresión de proteína (Normalizado)
Mimivirus de *A. polyphaga*	1.797210	0.8986	4.073458	0.9486
M. genitalium	1.844102	0.9221	4.041013	0.935
N. equitans	1.834140	0.9171	4.099902	0.9425
S. pombe (cr. 3)	1.928483	0.9642	4.165102	0.9637

Tabla 3: Comparación entre la compresión de las distintas secuencias utilizadas

Los valores de bps, tanto de la información genética codificante como de la cadena protéica traducida son altos. De hecho, son prácticamente todos son superiores a 0.9 cuando se normaliza, lo cual quiere decir que se acercan a la aleatoriedad completa, la cual tendría el valor máximo posible (es decir, la unidad, después de la normalización). Otro hecho notable es que en la proteína, la complejidad de las cadenas aumenta y la razón de compresión disminuye, lo cual en sí mismo indica ciertas particularidades tanto de las cadenas como de los códigos. ---discutir un poco qué implica que la complejidad de la cadena pueda aumentar, en función a la asimetría del código.

En el caso de las proteínas, estos valores elevados son aún más evidentes, pues existe un punto de referencia y comparación: los valores obtenidos de la compresión de las cadenas que nosotros traducimos utilizando los códigos generados con el programa de Netlogo. Debido a que todos los valores de la gráfica representan secuencias traducidas a partir de la misma cadena de DNA, la información que podemos obtener de estas figuras se refiere predominantemente a los códigos genéticos. La distribución de las secuencias generadas por todos estos códigos tiene una distribución regular; los datos estadísticos básicos se muestran en la tabla 4. La localización del valor de compresión del proteoma "real" (es decir, de aquel que fue traducido utilizando el código genético estándar) se encuentra consistentemente alejado del valor promedio, en la parte de la cola derecha de la gráfica y muy cercano al valor máximo de 4.322 bps. Podemos afirmar, pues, que el código genético es distinto a la mayoría de los códigos. Las secuencias de DNA de alta complejidad no necesariamente tienen que ser traducidas a una cadena de aminoácidos con una complejidad similar, como ya se ha discutido; no obstante, la estructura del código genético estándar logra conservar o incluso aumentar esta incompresibilidad. Nosotros

sugerimos que esta propiedad, al menos desde una visión puramente informacional, se debe sobre todo a la estructura de redundancia que presenta. Por supuesto, es necesario tomar en cuenta que el mensaje analizado y el código genético funcionan, dentro de la célula, complementándose el uno al otro, por lo cual la conservación de complejidad muy probablemente tenga raíces en la estructura del mensaje, el ordenamiento por familias del código, u otros factores además de la redundancia de cada uno de los aminoácidos. A continuación se presenta una tabla con los parámetros estadísicos de los datos obtenidos; es conveniente recordar que aunque la distribución se ordena normalmente, la escala es logarítmica, por lo que el incremento entre la complejidad promedio y la complejidad del proteoma es exponencial.

	Promedio de la población de proteomas artificiales	Código estándar	σ de la distribución de proteomas artificiales	Diferencia entre el promedio de proteomas artificiales y el valor del cod. st	Desviaciones estándar entre el promedio de proteomas artificiales y el cod st.
A.polyphaga mimivirus	3.8294	4.0410	0.2135	0.2441	1.1430
M.genitalium	3.7742	4.0999	0.1801	0.2668	1.4820
N.equitans	3.8377	4.0735	0.2131	0.2622	1.2300
S.pombe	3.9440	4.1651	0.1112	0.2211	1.9880

Tabla 4: Comparación estadística entre la distribución de proteomas generadas y el código genético estándar

En este punto de la discusión, creo conveniente señalar algunas reflexiones acerca de las distribuciones que interactúan en la representación de nuestros datos. Las gráficas que fueron presentadas en la sección de Resultados resumen la compresión de los datos de cadenas específicas. En este caso, las cadenas son los proteomas de cuatro organismos distintos, producidos por mil códigos artificiales generados por nosotros. Estos datos son sólo un subconjunto de los mensajes que se pueden producir a partir de los genomas utilizados, pero proveen una buena idea de la distribución del conjunto total. Sin embargo,

en la distribución de códigos resultantes hay una distribución implícita que se debe considerar para comprender uno de los factores que componen los resultados obtenidos.

Para ser traducido, cada uno de los genomas tuvo que pasar a través de una colección determinada de relaciones simbólica con distintas características: cada uno de los aminoácidos puede estar más o menos representado en el código, la proporción de los aminoácidos puede ser homogénea o algunos pueden estar sobrerrepresentados, puede existir redundancia en las familias de codones o entre codones cercanos, etc. Todas estas características definen la estructura de cada código. Cuando se considera que cada uno de los signos del código representa un objeto determinado, ya sea los tripletes de nucleótidos o los aminoácidos, los cuales tienen características físicas propias, es posible considerar más factores en la estructura del código, como hidrofobia e hidrofilia de los aminoácidos, tamaño de las moléculas o precursores químicos. Sin embargo, una característica sirve para agrupar al enorme conjunto de códigos posibles en unos cuantos subconjuntos de propiedades determinadas: el número de aminoácidos que están representados.

Como todos los códigos son generados al azar, es posible que no todos los aminoácidos sean asociados a algún codón. Por lo tanto, dentro del conjunto de códigos posibles existirá un subconjunto que cuente con los 20 aminoácidos, otros que sólo tengan reglas de traducción para 19, e incluso una minoría que haya asignado un mismo aminoácido a todos los codones (lo cual es, por supuesto, una improbabilidad estadística). Cada uno de estos subconjuntos tendrá un máximo de entropía informacional, determinado por el número de aminoácidos que utiliza. Por ejemplo, los códigos de 20 aminoácidos producen cadenas que tienen el ya discutido límite superior de 4.322 bps; si un aminoácido está ausente del código, las cadenas producidas sólo contarán con 19 símbolos y su valor máximo de entropía informacional es 4.248 bps. Los códigos de cuatro aminoácidos producirán cadenas similares a las de ácidos nucléicos, con límite superior de 2 bps. Por último, un código que traduzca cualquier codón a un mismo aminoácido producirá cadenas completamente repetitivas y, por lo tanto, su límite superior (e inferior) es 0 bps. El número total de códigos en cada clase también aumenta, como es predecible. Los códigos compuestos por un sólo aminoácido son veinte; los códigos posibles con 20 aminoácidos en 64 codones son aproximadamente 1.8447×10^{83}. La fórmula para obtener el número de códigos posibles es

$$(a^c)\ (t!\ /\ a!\ (t\text{-}a)!),$$

en donde *t* es el número total de aminoácidos (que siempre es 20), *a* es el número de aminoácidos que se están ocupando en ese conjunto de códigos, y *c* es el número de codones (que siempre es 64).

Si el conjunto de códigos que utilizan los veinte aminoácidos es el mayor y consecuentemente el más representado en nuestra simulación, ¿por qué la forma de la curva obtenida en nuestros resultados (figs. 5-8) no es igual a la forma de la curva del número de códigos? La primera tiene una distribución normal, cuya media se encuentra cercana al máximo de entropía informacional; la segunda es una curva de aumento exponencial, cuyo máximo se alcanza en el conjunto de los códigos que utilizan 20 aminoácidos. Para responder a esta pregunta es necesario recordar que las curvas que hemos presentado son las frecuencias de la complejidad de los proteomas. Esto quiere decir que dependen de las características de la cadena que ha sido traducida. Si una cadena de DNA tiene una complejidad mínima (como, en el caso extremo, las cadenas que sólo están compuestas por un solo nucleótido repetido) dará origen a cadenas de proteínas de muy poca complejidad, independientemente del código con el que se realice la traducción. Así, hay una interacción cercana entre el código genético y la cadena original de DNA para conservar, reducir o ,en algunos casos, incluso aumentar la complejidad de la cadena de aminoácidos.

Con estas consideraciones en mente, a continuación se analizan los resultados de nuestra simulación, y en particular las características de algunos de los códigos generados aleatoriamente. En la siguiente tabla se presentan los números de identificación de los códigos que produjeron las cadenas con mayor compresión y menor compresión para cada organismo. En la tercera columna se enlistan los números de los códigos que produjeron las cadenas de compresión más cercana al proteoma producido por el código estándar.

Organismo	Códigos de los proteomas de mayor complejidad	Códigos de los proteomas de menor complejidad	Códigos que produjeron proteomas con valores cercanos al del código estándar
Mimivirus de *Acanthamoeba p.*	805 (4.218) 713 (4.193) 404 (4.146)	794 (3.341) 374 (3.412) 79 (3.428)	260 (4.078) 987 (4.076) 242 (4.071)
Mycoplasma genitalium	805 (4.161) 404 (4.106) 761 (4.097)	794 (3.204) 162 (3.257) 16 (3.297)	563 (4.045) 340 (4.043) 678 (4.041)
Nanoarchaeum equitans	805 (4.194) 402 (4.169) 713 (4.145)	794 (3.321) 41 (3.432) 79 (3.442)	837 (4.105) 368 (4.100) 987 (4.098)
Schizosaccharomyces pombe	805 (4.239) 713 (4.212) 404 (4.186)	374 (3.521) 568 (3.530) 794 (3.547)	866 (4.163) 939 (4.162) 761 (4.162)

Tabla 5: Resumen de los códigos más importantes para cada organismo

En la tabla se puede observar que hay códigos que aparecen con más frecuencia. Por ejemplo, los códigos 805, 713 y 404 consistentemente producen cadenas de poca compresión (i.e., de complejidad elevada); de manera similar, los códigos 794, 374 y 79 aparecen frecuentemente en la columna de las cadenas más compresibles. En el caso de los códigos de compresión similar al código genético estándar únicamente hay una repetición (el del código 987 en Mimivirus y en *N. equitans*)

	U	C	A	G	
U	UUU C UUC C UUA W UUG M	UCU F UCC S UCA W UCG L	UAU N UAC E UAA A UAG Q	UGU T UGC L UGA M UGG Q	U C A G
C	CUU P CUC R CUA P CUG F	CCU H CCC A CCA T CCG D	CAU R CAC N CAA Q CAG I	CGU R CGC S CGA L CGG F	U C A G
A	AUU L AUC D AUA K AUG E	ACU T ACC T ACA A ACG H	AAU S AAC G AAA V AAG D	AGU I AGC W AGA A AGG W	U C A G
G	GUU N GUC Y GUA H GUG I	GCU L GCC E GCA Q GCG Y	GAU H GAC Y GAA Y GAG S	GGU R GGC T GGA E GGG D	U C A G

Tabla 6: Código no. 805 (Compresión mínima)

	U	C	A	G	
U	UUU T UUC W UUA M UUG M	UCU C UCC N UCA F UCG E	UAU T UAC F UAA R UAG R	UGU G UGC D UGA F UGG T	U C A G
C	CUU T CUC V CUA K CUG I	CCU S CCC E CCA K CCG C	CAU F CAC F CAA W CAG F	CGU F CGC S CGA G CGG P	U C A G
A	AUU T AUC M AUA F AUG Q	ACU Y ACC T ACA E ACG N	AAU Y AAC D AAA F AAG F	AGU A AGC H AGA F AGG Y	U C A G
G	GUU N GUC D GUA N GUG W	GCU T GCC T GCA F GCG V	GAU F GAC D GAA F GAG T	GGU G GGC T GGA T GGG N	U C A G

Tabla 7: Código no. 794 (Compresión máxima)

	U		C		A		G		
U	UUU	V	UCU	W	UAU	S	UGU	Y	U
	UUC	Y	UCC	R	UAC	F	UGC	F	C
	UUA	Y	UCA	R	UAA	D	UGA	D	A
	UUG	P	UCG	E	UAG	W	UGG	W	G
C	CUU	H	CCU	S	CAU	E	CGU	K	U
	CUC	M	CCC	N	CAC	W	CGC	R	C
	CUA	I	CCA	M	CAA	C	CGA	R	A
	CUG	T	CCG	D	CAG	M	CGG	R	G
A	AUU	S	ACU	C	AAU	P	AGU	M	U
	AUC	K	ACC	A	AAC	K	AGC	T	C
	AUA	A	ACA	F	AAA	T	AGA	F	A
	AUG	Q	ACG	A	AAG	V	AGG	D	G
G	GUU	E	GCU	N	GAU	L	GGU	P	U
	GUC	D	GCC	W	GAC	T	GGC	I	C
	GUA	R	GCA	H	GAA	L	GGA	W	A
	GUG	R	GCG	T	GAG	N	GGG	R	G

Tabla 8: Código no. 987 (Compresión similar al código estándar)

	U		C		A		G		
U	UUU	G	UCU	S	UAU	H	UGU	T	U
	UUC	N	UCC	W	UAC	A	UGC	S	C
	UUA	E	UCA	T	UAA	Q	UGA	E	A
	UUG	A	UCG	K	UAG	G	UGG	I	G
C	CUU	H	CCU	V	CAU	F	CGU	I	U
	CUC	P	CCC	Y	CAC	H	CGC	F	C
	CUA	I	CCA	H	CAA	I	CGA	*	A
	CUG	N	CCG	F	CAG	Y	CGG	P	G
A	AUU	W	ACU	G	AAU	N	AGU	P	U
	AUC	I	ACC	M	AAC	Q	AGC	*	C
	AUA	I	ACA	F	AAA	Y	AGA	I	A
	AUG	W	ACG	V	AAG	F	AGG	P	G
G	GUU	F	GCU	V	GAU	K	GGU	Y	U
	GUC	N	GCC	Y	GAC	K	GGC	N	C
	GUA	D	GCA	*	GAA	Y	GGA	K	A
	GUG	F	GCG	L	GAG	P	GGG	Q	G

Tabla 9: Tabla 8: Código no. 678 (Compresión similar al código estándar)

	U		C		A		G		
U	UUU	E	UCU	I	UAU	K	UGU	K	U
	UUC	P	UCC	I	UAC	G	UGC	P	C
	UUA	C	UCA	Y	UAA	A	UGA	*	A
	UUG	P	UCG	E	UAG	P	UGG	P	G
C	CUU	*	CCU	T	CAU	K	CGU	S	U
	CUC	A	CCC	G	CAC	P	CGC	D	C
	CUA	M	CCA	K	CAA	C	CGA	I	A
	CUG	Y	CCG	P	CAG	Y	CGG	M	G
A	AUU	Y	ACU	S	AAU	R	AGU	H	U
	AUC	S	ACC	P	AAC	R	AGC	T	C
	AUA	L	ACA	*	AAA	W	AGA	K	A
	AUG	M	ACG	P	AAG	P	AGG	L	G
G	GUU	T	GCU	V	GAU	K	GGU	C	U
	GUC	Q	GCC	N	GAC	L	GGC	F	C
	GUA	D	GCA	K	GAA	T	GGA	D	A
	GUG	N	GCG	A	GAG	N	GGG	H	G

Tabla 10: Tabla 8: Código no. 866 (Compresión similar al código estándar)

Compresión mínima (805)	Compresión máxima (794)	Estándar	Similar a estándar (987)	Similar a estándar (678)	Similar a estándar (866)
K=1	K=2	K=2	K=3	K=2	K=2
N=3	N=5	N=2	N=3	N=1	N=4
R=4	R=3	R=6	R=8	R=8	R=3
S=5	S=2	S=6	S=3	S=7	S=3
T=5	T=11	T=4	T=5	T=2	T=1
I=3	I=1	I=3	I=3	I=6	I=3
M=2	M=3	M=1	M=4	M=3	M=5
Q=4	Q=2	Q=2	Q=2	Q=4	Q=5
H=4	H=2	H=2	H=2	H=6	H=2
P=3	P=2	P=4	P=3	P=2	P=4
L=5	L=1	L=6	L=4	L=4	L=2
E=6	E=3	E=2	E=3	E=2	E=6
A=5	A=2	A=4	A=5	A=4	A=3
D=5	D=4	D=2	D=6	D=2	D=7
G=3	G=3	G=4	G=1	G=3	G=6
V=1	V=4	V=4	V=2	V=2	V=4
Y=6	Y=5	Y=2	Y=4	Y=3	Y=6
W=5	W=4	W=1	W=7	W=8	W=3
C=3	C=3	C=2	C=3	C=3	C=4
F=4	F=15	F=2	F=5	F=6	F=4

Tabla 11: Redundancias de los códigos presentados anteriormente

Como era esperado, en el código que produce cadenas con más posibilidad de compresión existen aminoácidos mucho más representados; en este caso particular, F, con 15 codones asignados, y T con 11. Por su parte, el código genético que produce cadenas menos compresibles (el 805), los aminoácidos están distribuidos de manera regular en el código, por lo que su aparición tiende a ser equiprobable y por lo tanto se acerca a una secuencia completamente aleatoria. También es posible notar que los códigos parecidos no dependen de una similaridad con la redundancia específica de cada aminoácido del código estándar: por ejemplo, el triptofano, que en el código estándar sólo cuenta con un codón, en los códigos que arrojan compresiones similares tiene una redundancia variables (de 7, 8 y 3, en los códigos presentados en este escrito). Sin embargo, para comprender un poco más estos resultados es necesario analizar las conexiones entre las medidas que estamos utilizando y algunos fenómenos físicos, en particular la entropía.

Conexiones con las ciencias físicas

La información, tal como es definida por Shannon (1949), la complejidad algorítmica y la representación simbólica de los fenómenos físicos tienen raíces profundas en la termodinámica. No es una casualidad que la medida de información de Shannon sea llamada "entropía"; tampoco es una casualidad que la fórmula central de la teoría informacional copie de manera casi exacta las fórmulas de Boltzmann y Gibbs. A partir de este hecho básico, las conexiones y analogías se multiplican; por ejemplo, el aumento de entropía informacional debido a cualquier proceso implica una incapacidad de reversión al estado original; además, la entropía informacional tiende a incrementar con el paso del mensaje a través de un canal si no se establecen restricciones que lo eviten. La conexión fundamental entre procesos informacionales proveyó la respuesta de Leó Szilárd a la famosa paradoja del demonio de Maxwell: el demonio no viola la segunda ley de la termodinámica, pues la información que utiliza para discriminar las partículas puede ser convertida en energía (ver Toyabe et al, 2010, en donde aparentemente también se demuestra experimentalmente la hipótesis de Szilárd). Dada la insistencia que hemos mostrado a lo largo de este trabajo acerca de la importancia del funcionamiento simbólico

de los seres vivos, se justificará la necesidad de exponer algunas consideraciones y argumentos acerca de las características de la información biológica.

Las expresiones de entropía informacional (H) y entropía termodinámica (S) son exactamente las mismas; ambas se refieren a la interpretación del concepto de entropía como una cantidad que se relaciona directamente con el número de microestados que puede tener un macroestado (la cual, a su vez, es la interpretación que dio origen a la concepción vulgar, pero en cierto sentido incompleta, de que la entropía es una medida del desorden). La diferencia entre microestados y macroestados se puede explicar fácilmente pensando en un sistema consistente en un recipiente cerrado lleno de gas. En la teoría de la mecánica estadística, es imposible una descripción completa de este sistema. Todos sus elementos, como moléculas y átomos, están en constante movimiento y sus posiciones y velocidades se modifican continuamente, por lo que en cada momento surge una configuración distinta de estas partículas; cada una de estas configuraciones es un *microestado.* Sin embargo, sus características macroscópicas, como la temperatura y la presión, se mantienen constantes mientras el sistema se encuentre cerrado. Por lo tanto, un mismo macroestado, definido por estas propiedades observables macroscópicamente, puede tener un número enorme de microestados distintos. No es necesario conocer el microestado preciso en el momento de la observación (es decir, los lugares que ocupan las moléculas individuales, o las velocidades que cada una tiene) para poder describir el macroestado del sistema mediante magnitudes fácilmente medibles.

Ludwig Boltzmann fue el primero que intentó describir la relación estadística entre microestados y macroestados. La fórmula original de Boltzmann (la cual se puede observar en su lápida funeraria) es la siguiente:

$$S = k \log W,$$

que iguala a la entropía S al producto del logaritmo de W (el número de microestados) y k (la constante de Boltzmann que define las unidades de la entropía e incluye el número de Avogadro). Sin embargo, esta fórmula sólo describe un caso aislado de la mecánica estadística: en el que todos los microestados son equiprobables, lo cual ocurre únicamente en un gas ideal. Las predicciones de la fórmula son usualmente erróneas en los gases reales,

debido a las diferencias de probabilidades del sistema para encontrarse en un microestado o en otro. Por lo tanto, Gibbs generaliza la fórmula a

$$S = - k \Sigma \, pi \log pi,$$

en la que *pí* es igual a la probabilidad de aparición de cada uno de los microestados (1,2,3...i). Por su parte, la fórmula de la entropía informacional de Shannon es

$$H = - k \Sigma \, pi \log pi,$$

En ella, la constante k se vuelve irrelevante pues sólo funciona como elección de las unidades del valor de entropía, las cuales generalmente son bits. Como se puede observar, es prácticamente idéntica a la fórmula de la entropía de la mecánica estadística; aunque Shannon llegó a ella independientemente, es evidente que se dio cuenta de la relación entre ambas fórmulas. No sólo comenta la igualdad en su artículo clásico (Shannon, 1948), sino que nombra a su medida con un término termodinámico.

La conexión entre la entropía informacional y la entropía de Boltzmann-Gibbs se encuentra en el proceso probabilístico que rige el cambio entre estados. La base matemática de ambos se encuentra en los procesos de Markov, los cuales son una expresión que dicta la probabilidad de que un estado determinado cambie a otro.

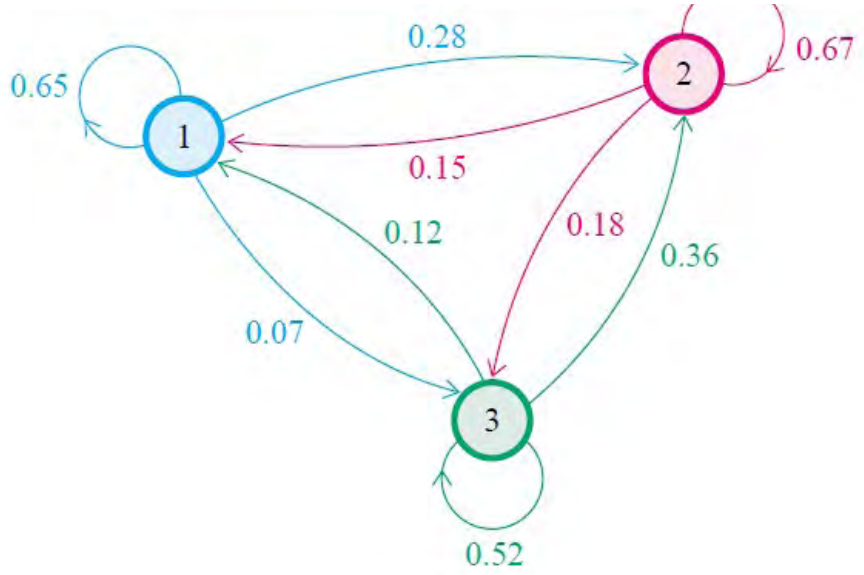

Fig. 10: Una cadena de Markov, especificando la probabilidad de pasar de uno de los estados a otro (tomado de Greenwell et al, 2003)

En el caso de la mecánica estadística, el proceso de Markov determina la posibilidad de pasar de un microestado a otro del mismo macroestado. Cuando todos los microestados son equiprobables e independientes entre sí, la fórmula se simplifica en la forma original de Boltzmann. En el caso de la teoría informacional, los microestados y los macroestados pueden tener distintas identidades, dependiendo del nivel de descripción del sistema que se elija.

Como lo señalan Nicolis y Prigogine (1989), la corrección de Gibbs a la fórmula original de Boltzmann tiene algunas implicaciones fundamentales para entender la dinámica del universo físico, y de la evolución biológica. Si todos los microestados de un macroestado fueran equiprobables, el cambio entre uno y otro sería indiferente y no existiría ningún tipo de incremento de complejidad. En el caso de los sistemas físicos, los estados estables alejados del equilibrio tienen una probabilidad alta de surgir si existe la suficiente entrada de energía desde el exterior del sistema. La autoorganización que deriva de ellos no es, por lo tanto, una fenómeno excepcional que sea difícil de alcanzar; es un estado entre muchos otros posibles del sistema, pero que tiene una probabilidad elevada. Lo

mismo ocurre en el caso de los seres vivos y del caso análogo de la entropía informacional: Nicolis y Prigogine señalan que el número de cadenas simbólicas posibles es enorme, pero algunos factores o características propias del sistema (como las restricciones impuestas por la naturaleza fisicoquímica de los bloques que las componen, la selección de cadenas con contenido semántico, etc) favorecen ciertas configuraciones[7]. Así, el espacio de secuencias viables que se tiene que explorar para encontrar puntos óptimos o soluciones funcionales se reduce considerablemente. En ambos casos (la preferencia de un sistema físico por ciertos estados fisicoquímicos o el favorecimiento de un conjunto reducido de secuencias de biopolímeros), ocurre una ruptura de simetría. La presencia de una estructura asimétrica del sistema es uno de los dos requisitos para el surgimiento de la información; el otro es la impredecibilidad expresada por la entropía de Shannon.

La diferenciación de microestados y macroestados implica un cambio en el nivel de descripción del sistema. Esta característica es, según el texto ya citado de Nicolis y Prigogine, la huella más importante de la complejidad y en los sistemas informacionales se presenta de manera múltiple. Para entender esta presencia polifacética, consideremos las diversas identidades que pueden adquirir los conceptos abstractos de "estado", importados de la teoría física de la mecánica estadística, en diversas manifestaciones informativas de los sistemas biológicos.

1) Cuando se resuelve la ecuación de Shannon, *pi* contabiliza la probabilidad de aparición de cada símbolo en un lugar determinado de la secuencia. Así, i = 1, 2, 3... n describen cada uno de los microestados del sistema. Por lo tanto, los microestados representan a cada una de las letras del alfabeto que compone a la secuencia. Por ejemplo, en el caso del DNA, los microestados son cuatro: A, G, C y T; en el caso de los polipéptidos, los microestados posibles son 20. El macroestado, representado por cada uno de los espacios de la secuencia, es irrelevante: la descripción de cada uno de ellos no es más que un lugar que puede ser ocupado por un símbolo.
2) Cada valor de entropía de Shannon define un conjunto de secuencias; éste es un nivel descriptivo más alto. Cuando la longitud de la secuencia se mantiene

7 E. Hernández, comunicación personal

constante, el conjunto está compuesto por sus posibles permutaciones. En este punto, se agrega una dimensión temporal al espacio de estados del sistema: la secuencia de símbolos puede ser leída como una serie de tiempo obtenida mediante la observación periódica de un sistema físico. Los microestados son, entonces, cada una de las series de tiempo posibles que conserven el mismo número de apariciones de los estados. Dicho de otra manera: el macroestado es descrito únicamente por el valor de la entropía informacional; los microestados son cada una de las cadenas que cumplen con las probabilidades de aparición de cada uno de los símbolos.

3) Existe otro nivel de descripción que aparece cuando la información se procesa a través de un código. La traducción como la lleva a cabo el código estándar, según se expondrá en una sección posterior, implica un proceso irreversible en el que se pierde información y aumenta la entropía informacional (Yockey, 2005). La colección de relaciones simbólicas que forma el código puede ser tomada como un espacio de estados independiente. Para ejemplificar, considérese el código genético estándar, el cual tiene una estructura de redundancia característica, en la que cada aminoácido está representado por un número definido de codones. Si se conserva la estructura de redundancia, pero se reasigna cada aminoácido a un codón distinto, los dos códigos representarían microestados del mismo macroestado. Si tomáramos una secuencia completamente aleatoria de DNA y la tradujéramos con cada uno de los códigos, las cadenas de aminoácidos resultantes tendrían complejidad y entropía informacional muy cercanas, aunque la secuencia como tal fuera muy distinta. Por otro lado, es fácil imaginar el resultado de un código que traduce todos los codones a un mismo aminoácido: en él, la redundancia es máxima y el alfabeto es reducido (de únicamente un símbolo, menor incluso que el alfabeto original de cuatro nucleótidos). Así, las cadenas producidas tendrán necesariamente una entropía informacional de cero, sin importar la entropía de la cadena original.
   Los códigos son, pues, un generador de entropía, y la traducción es un proceso que puede ser considerado un nivel más de descripción. Las relaciones entre la traducción y la mecánica estadística son un poco más oscuras que en los niveles anteriores, pero son innegables. Las distintas configuraciones de la tabla de los codones, así como las propiedades específicas de las relaciones simbólicas son los

microestados de un macroestado que puede ser definido con mediciones simples de diversas características, como la redundancia, el tamaño del alfabeto de aminoácidos y la entropía máxima del código. Por supuesto, es necesario llevar a cabo una investigación detallada acerca de la multiplicidad posible de niveles en sistemas físicos para realizar una analogía más concreta y explorar mutuamente ambos tipos de propiedades emergentes.

Tal vez parezca irrelevante discutir con tanta profundidad la teoría de Shannon, ya que en secciones anteriores se ha afirmado que no se utiliza directamente como la medida principal del presente trabajo. Es cierto: la entropía de Shannon resulta insuficiente para los fines que perseguimos en esta investigación. Sin embargo, la medida informacional que más se acerca a las medidas que utilizamos (i.e., la complejidad de Kolmogorov) no es una medida independiente de la entropía informacional. Todos los teóricos que trabajaron para proponerla (Solomonoff, Kolmogorov, Chaitin) intentaban conscientemente llenar una laguna teórica de la teoría matemática de la información. Mientras que Shannon se ocupaba de las características de la fuente que producía a los símbolos, sin tomar en cuenta las secuencias originales, Kolmogorov observaba directamente al mensaje individual, sin importarle cómo se originó. Existen ciertas conexiones entre ambas, pero son limitadas: Grünwald y Vitányi (2004) exponen un teorema que demuestra que en complejidades bajas, las medidas de Kolmogorov y de Shannon convergen, pero que en complejidades altas (como la de la información biológica que este trabajo analiza) pueden tener valores muy distintos.[8] Como ya se expuso en la discusión metodológica, en el presente trabajo consideramos las dos visiones distintas de complejidad y entropía y proponemos un método que, aunque carece de la elegancia de las abstracciones totales de Shannon y Kolmogorov, es una manera repetible de transportar diversos puntos centrales de ambas teorías a un

8 Aunque los autores aclaran las repercusiones conceptuales de dicha información, la interpretación matemática que puedo realizar de dicho teorema es limitada, en el mejor de los casos. Sin embargo, me parece pertinente citarlo literalmente (p.15):

Theorem 2.10 *Let f be a computable probability mass function (Section 1.2) $f(x) = P(X = x)$ on sample space $\mathcal{X} = \{0,1\}^*$ associated with a random source X and entropy $H(X) = \sum_x f(x) \log 1/f(x)$. Then,*

$$0 \leq \left(\sum_x f(x)K(x) - H(X) \right) \leq K(f) + O(1).$$

mismo terreno. Así, intentamos crear una visión informacional que pueda ser aplicada a diversos niveles de descripción física.

Implicaciones biológicas

Distintos tipos de semiótica biológica

Hasta ahora, los resultados presentados se han discutido de manera más o menos abstracta, desde el punto de vista de la teoría de la información y de las relaciones con ciertos parámetros físicos, como la entropía, los estados del sistema o la transmisión de información. Es necesario analizar ahora algunas de las implicaciones más concretas y centradas en los sistemas biológicos, que finalmente son nuestro interés central. Los estudios biosemióticos son una manera diferente de observar fenómenos que ya se conocen desde hace décadas; por tal razón, espero que las consideraciones siguientes arrojen una luz distinta sobre algunos procesos, y que funcionen como un apoyo y reforzamiento de algunas ideas que varios autores han propuesto acerca de los seres vivos y sus características.

Para los sistemas biológicos, los comportamientos semióticos son fundamentales. Se encuentran, por ejemplo, en el conjunto de relaciones abstractas que constituyen el código genético, que no sólo es la manera que tienen los seres vivos para procesar información, sino que es una de las características únicas de los seres vivos que los distinguen del resto del mundo. Como se ha desarrollado en escritos anteriores (Mercado 2009), no podemos encontrar el mismo tipo de manejo de información o de codificación espontánea en otro tipo de sistemas físicos, sin importar la complejidad de éstos. El código genético, como sistema semiótico fundamental, es la causa primordial de la separación del fenotipo y el genotipo, la cual se encuentra en la base de la evolución darwiniana[9]. Otros procesos

9 De esta manera, el uso de códigos da cuenta de una de las definiciones de vida, la posibilidad de llevar a cabo evolución por selección natural. Otro punto de las definiciones tradicionales, la compartamentalización y la separación del medio, también puede incluirse como un funcionamiento semiótico, pues la semiótica causa una separación y reconocimiento por parte de los organismos de un cuerpo propio (*Innenwelt*) y el mundo exteno (*Umwelt*) (von Uexküll, 1982). Sin embargo, estas consideraciones exceden en complejidad y alcance a los objetivos del presente trabajo.

posibles de evolución, incluyendo las propuestas epigenéticas y la evolución neutral de Kimura (1985), contienen elementos semióticos, tanto del código como de la manifestación funcional de los genes. Además del código genético, Barbieri (2003) identifica códigos en otros niveles, como en el procesamiento de las cadenas de RNA por *splicing* o la señalización celular, que toman un papel fundamental en el funcionamiento integral de los sistemas vivos. Sin embargo, todos estos fenómenos semióticos difieren de los tradicionalmente estudiados en que no implican ningún tipo de interpretación ni de idealización del objeto simbolizado (lo cual es lógico, pues no existe una mente que realice el proceso). Es por eso que Sharov (2009, 2010) llama a esta dinámica "semiosis vegetativa", compuesta de protoiconos, protoíndices y protosímbolos, la cual sirve como base para procesos semióticos más complejos, que incluyen clasificación, modelación y hermenéutica. Esta afirmación es complicada de sostener, particularmente porque es difícil establecer una línea inequívoca entre la semiosis vegetativa y la semiosis "tradicional". Sin embargo, creo que existen ciertas implicaciones que se reflejan en este trabajo. Por ejemplo, en el escenario hipotético, mencionado en la sección "El código genético", sugerimos que la conformación de un código robusto, que pueda interpretar una variedad amplia de mensajes, necesita de la interacción social de varios sistemas codificantes. Hay diversas posibilidades para dicha interacción; Woese y sus colaboradores (Woese et al, 2000; Vetsigian et al 2006; Goldenfeld y Woese, 2011) sugieren que se puede llevar a cabo con un mecanismo análogo a la transmisión horizontal; es también posible que sea un comportamiento esperado debido a la retroalimentación constante entre código y mensaje. De cualquier manera, la semiótica básica funciona como un punto de partida para estructurar a los sistemas semióticos más complejos, pues los niveles "vegetativos" y los superiores se afectan mutuamente.

Retroalimentación

Hemos mencionado que el código genético, como prácticamente cualquier aspecto de los sistemas vivos, fue estructurado bajo la infuencia de varios factores externos, y a su vez fue la causa de modificación de otros sistemas. Se ha afirmado, por ejemplo, que el código y los mensajes que éste traduce se afectan mutuamente, y también que el código, como nivel

básico de la semiótica, tiene efectos en el nivel social, el cual a su vez modifica la estructura del código. Tal vez estas consideraciones puedan parecer triviales, pero creo necesario hacer énfasis en ellas. La visión contraria se deriva de un pensamiento tanto atomista (en el que las partes de los sistemas biológicos son elementos aislados) como reduccionista (que pretende explicar el funcionamiento de todo el sistema únicamente mediante el análisis de sus componentes mínimos), y permea, aún en nuestros días, la concepción de la naturaleza y funcionamiento de los seres vivos. No es infrecuente que se piense que un gen codifica para una característica discreta y definida del cuerpo de un organismo, y que el fenotipo puede ser mapeado, punto por punto, a distintas regiones del genoma. Una versión menos caricaturizada pero más común de esta percepción, es la representación de los organismos como una serie de cascadas de señalización molecular, en la que prácticamente cada proceso se origina y termina en una instrucción genética.

Para intentar acercarnos a esta cuestión desde un enfoque de la biología de sistemas, en este escrito discutimos algunas de las influencias externas que pudieron afectar y ser afectadas por el surgimiento del código, aunque debido al enfoque de la investigación, dicha discusión será necesariamente superficial e incompleta. Como únicamente tratamos aspectos informacionales y, más precisamente, de traducción y procesamiento de la información, necesariamente tenemos que obviar gran parte de las caractísticas fisicoquímicas de los elementos. Además, aunque hay muchas propuestas de los escenarios en los que se llevó a cabo el surgimiento y evolución temprana de la vida, es difícil afirmar con certeza las condiciones en las que estos procesos ocurrieron. El ambiente del origen de la vida sigue siendo, en el presente momento, una pregunta abierta.

Dejando un poco de lado las propiedades físicas de los elementos informacionales, nuestra discusión girará en torno a dos consideraciones: la influencia que pudo tener el código para estructurar los mensajes biológicos, y la influencia de los mensajes en la estructuración del código. El primer caso (el efecto del código en el mensaje) resulta relativamente sencillo. El producto protéico de un mensaje de DNA depende directamente de las asignaciones de los distintos codones a los aminoácidos.Una vez que se ha establecido cierto número de relaciones simbólicas, los nuevos mensajes que surjan deben poder ser traducidos a proteínas funcionales en ese contexto semiótico. La aparición de nuevos aminoácidos en el código también tiene un efecto sobre la información genética,

haciendo que codones que previamente se asociaban con otro aminoácido o que significaban una interrupción en traducción tengan otra identidad; además, un nuevo aminoácido aumenta la complejidad máxima de los mensajes traducidos, y permite introducir elementos con propiedades distintas, y por tanto posibles funciones nuevas, a las proteínas. El código también afecta al mensaje de manera post-traduccional, pues un código con más redundancia tenderá a producir proteínas de menor complejidad.

Aunque los mecanismos son menos evidentes, hay varias maneras posibles en las que el mensaje pueda afectar al código. Por ejemplo, Bedian (1982, 2001) muestra en un modelo muy simple cómo puede favorecer una asignación simbólica. Especificando algunas posiciones funcionales en la proteína que deben ser ocupadas por determinados aminoácidos, ciertos códigos tienen un mejor desempeño; esto ocasiona una ruptura de simetría en los puntos de estabilidad en el código. . En el escenario hipotético del origen del código presentado anteriormente, se proponen dos posibles efectos de los mensajes sobre el código. En primer lugar, la necesidad de conservar características informacionales de los mensajes genéticos para permanecer en el rango de complejidad funcional, como lo proponen Abel y Trevors (2005), ocasiona que exista sólo un subconjunto de códigos viables de los millones posibles. En segundo lugar, la presencia de varios agentes independientes, y por lo tanto, de varios tipos de mensaje, hace necesario un código robusto que sea lo suficientemente flexible como para poder interpretar mensajes de distintos agentes, surgido mediante consenso. [10]

Debido al efecto mutuo entre código y mensaje, es posible hablar de una coevolución entre ellos. El surgimiento de nuevas características en el código hace que los mensajes ya presentes sean traducidos de manera distinta, y que los mensajes nuevos se produzcan en un contexto distinto; la necesidad de conservar características informacionales y la complejidad de los mensajes ocasiona que el código tenga restricciones y reduce el número de colecciones viables de relaciones semióticas. Sin embargo, es conveniente tomar en cuenta el planteamiento de Francis Crick (1968) en su hipótesis del accidente congelado: después de un periodo de flexibilidad, el código necesariamente se estabiliza en un punto probablemente subóptimo, pues los mensajes que

10 Un modelo de surgimiento de códigos consenso entre una población de agentes muestra que dichos puntos óptimos de traducción son, la mayoría de las veces, múltiples, y que no hay necesariamente un único código viable (datos personales sin publicar)

dependen de él son importantes para el mantenimiento del sistema, y cualquier cambio en la traducción traería consecuencias letales para el organismo. Osawa y Jukes (Osawa et al, 1992) proponen como mecanismo de cambio tardío en la asignación de codones una especie de cuello de botella evolutivo, en el que el genoma es tan pequeño y un codón determinado aparece tan pocas veces que puede ser modificado sin afectar demasiado al organismo en cuestión. Los autores proponen este proceso como causa de los códigos genéticos alternativos que tienen algunos procariontes y organelos eucariontes, pero puede ser considerado como un posible mecanismo para las etapas tardías del establecimiento del código actual.

Asimetría del código y complejidad de la información.

Probablemente el resultado central del modelo presentado en este trabajo sea el elevado nivel de complejidad informacional de las secuencias biológicas. Tal característica no sólo se encuentra en las cadenas originales de DNA, sino que es conservado por el código genético estándar al traducirlas. Abel y Trevors (2005) afirman que las secuencias biológicas funcionales se encuentran más cercanas a la aleatoriedad que al orden, y que su complejidad es alta; aunque existen motivos y grupos de símbolos que tienen cierta correlación y tienden a aparecer juntos (por ejemplo, grupos funcionales pequeños de aminoácidos, regiones de hidrofobia en las que la proteína atraviesa la membrana), en general, las secuencias funcionales contienen un nivel alto de incertidumbre. Los autores argumentan que es un requisito para que los sistemas biológicos puedan evolucionar por medio de selección natural. Si las secuencias funcionales fueran completamente caóticas (es decir, se movieran en el terreno del azar puro) no serían informacionales; si estuvieran predeterminadas por leyes físicas rígidas (es decir, determinadas por necesidad en el sentido de Monod (1973), no habría variación posible ni flexibilidad, por lo cual no habría materia prima para la evolución darwiniana. Abel y Trevors no proveen ejemplos concretos de dichas secuencias; tampoco proponen un método unívoco de medición de la "complejidad funcional" [11] El acercamiento que proponemos en el presente trabajo es un

11 Posteriormente, Abel y Trevors colaboraron en un trabajo que intentaba cuantificar la "funcionalidad" de la información mediante la medición de sitios conservados entre genes homólogos (Durston et al, 2007). Por supuesto, esto es criticable, pues la funcionalidad no necesariamente implica una estasis informacional o un sitio activo particular. Esta técnica puede ser incluso contradictoria a las propuestas que habían realizado en

método nuevo para probar las predicciones de Abel y Trevors y una confirmación de su propuesta.

El elevado nivel de complejidad no se limita al DNA. Las proteínas son, de igual manera, secuencias de símbolos con función, por lo que la teoría de Abel y Trevors se sostiene en ellas. Sin embargo, como ya hemos mencionado algunas veces en este trabajo, la entropía máxima de las proteínas es de 4.322 bps debido al tamaño de su alfabeto, mientras que el DNA tiene una entropía máxima de 2 bps. Por ello, cuando mencionamos que la complejidad alta se conserva entre el DNA y la proteína codificada, se debe entender que se conserva su complejidad *relativa*; la complejidad absoluta asciende en varios órdenes de magnitud, al menos al ser medida con nuestro método. Esto se observa en la tabla 3.Sin embargo, como lo indica Yockey (2005) y contrariamente a lo que se podría pensar, este aumento de complejidad acarrea una disipación de energía y una pérdida de información. Como ya se ha mencionado en el apartado anterior, esta pérdida está directamente causada por la asimetría del código genético, la cual se presenta particularmente en la irreversibilidad de la codificación. El código genético es unidireccional porque, como todo código, depende de un mapeo único de las letras del alfabeto 1 (es decir, de los 64 codones del RNA) a letras del alfabeto 2 (los 20 aminoácidos). Dado que el código está casi saturado y prácticamente todas las letras del alfabeto de RNA tienen un equivalente en aminoácido, es necesariamente redundante. Por tal razón, dada una cadena de proteína es imposible saber cuál de los codones redundantes ha dado origen a cada uno de los aminoácidos.

Aunque Yockey lo descarte, el caso contrario no es imposible; es decir, que una letra del alfabeto de origen tenga varios significados y sea *ambigua*. En el código genético, la ambigüedad implicaría que una letra del alfabeto de RNA codifique para varios aminoácidos posibles. Esto no ocurre en los seres vivos, lo cual representa una asimetría más en el código: éste es redundante, mas no ambiguo. Es válido preguntarse si la ambigüedad tendría ventajas, y cómo sería la dinámica codificante en ese escenario. Si una misma cadena de DNA tuviera más de un significado, habría mucha más flexibilidad en los mensajes comunicados, y el espacio de exploración fenotípica sería mucho mayor. Además, la variación poblacional se incrementaría, por lo que los grupos tendrían, dentro de cierto

artículos previos.

rango, mayor resistencia a condiciones ambientales cambiantes. Sin embargo, las desventajas compensan con creces estas ventajas relativas. El único método para controlar la información ambigua es el análisis del contexto en el que se presenta. Como ejemplifica Barbieri (2003), una misma señal (acetilcolina) codifica para respuestas muy variadas e incluso opuestas en distintos tipos celulares, como músculo esquelético o músculo cardiaco. Sin embargo, para que esto ocurra, existe una maquinaria de adaptadores, dependientes del contexto celular, que interpretan el mismo mensaje de maneras distintas. Esto, por supuesto, implica un gasto energético mayúsculo. No obstante, la desventaja más evidente sería la poca confiabilidad de un proceso ambiguo; habría poco control en la producción de proteínas, cuya precisión es necesaria para llevar a cabo procesos indispensables en los organismos. La posibilidad de que un sígno tenga distintas relaciones simbólicas acarrearía, muy probablemente, efectos letales en la gran mayoría de los casos. Por lo tanto, sólo la redundancia tiene cabida en el código genético, no la ambigüedad.

El azar y la necesidad

Finalmente, creo necesario discutir brevemente los conceptos de arbitrariedad y aleatoriedad, que han aparecido frecuentemente a lo largo de este trabajo, y que están estrechamente relacionados entre sí. Dichos conceptos se mencionan en dos contextos: por un lado, en las técnicas empleadas para elaborar el modelo que presentamos; por el otro, en la discusión acerca de los posibles factores que han dado origen al código genético estándar. La arbitrariedad del código se presenta en la ausencia de necesidad física que inherentemente tienen los sistemas puramente simbólicos; la aleatoriedad, por su parte, se permea cada proceso físico en mayor o menor grado, y es parte integral de ellos. Que ambos estén presentes no implica que no haya una estructuración en el mundo físico: existen leyes y propiedades de cada una de las partes del sistema que lo proveen de tendencias, y, por tanto, de patrones, como las regularidades presentes en el ordenamiento de aminoácidos por su hidrofilia o hidrofobia, o su ruta de biosíntesis.

Este balance entre aleatoriedad, arbitrariedad y determinismo es visible en todas las escalas del descripción física. Es, por ejemplo, una de las razones que hicieron necesarias el surgimiento de la mecánica estadística. Es prácticamente imposible describir todas las

trayectorias de las partes de un sistema físico, como un gas en un recipiente cerrado. Sin embargo, la interaccion de dichas partes hacen que el sistema como un todo tenga tendencias predecibles en un nivel macroscópico. En los seres vivos, este tipo de efectos se multiplican tanto en número como en intensidad. Muchas veces, los procesos biológicos son caóticos, lo que implica que cambios pequeños tienen efectos desproporcionados. Además, y de manera más relevante, los sistemas biológicos están sometidos constantemente a contingencias históricas, lo que los hace adquirir características impredecibles y tomar caminos evolutivos insospechados.

En el código genético dichas contingencias pueden ser observadas fácilmente. A pesar de los patrones discutidos arriba, surgen varias preguntas: ¿por qué tiene precisamente ese grado de redundancia? ¿Qué efectos pudo tener las leyes de bamboleo (Crick, 1966) en la estructuración temprana del código? ¿Por qué el número de aminoácidos esenciales es veinte? La respuesta a estas preguntas radica en el proceso que Crick (1968) llamó "un accidente congelado": después de una etapa inicial en la que tomaron parte igualmente importante las afinidades fisicoquímicas, la facilitación de reacciones de síntesis de moléculas y las contingencias históricas, el código se tuvo que estabilizar en un punto en el que algún cambio de magnitud significativa traería consecuencias fatales para la organización del sistema. Es decir, como todo proceso biológico, el origen del código se puede explicar por la convergencia del azar, la necesidad *y* las eventuales restricciones.

Esta triada de causalidad sigue operando, por supuesto, hasta nuestros días. Se conocen casos de códigos genéticos alternativos, en el que un codón (frecuentemente el de inicio o los de término) tienen una asociación simbólica con un aminoácido distinto al código estándar, e incluso con aminoácidos no esenciales. También se sabe (((citas))) que cada grupo taxonómico hace uso de un dialecto particular de codones, y utiliza preferentemente alguna de las opciones de redundancia. Ésta es la razón por la que decidimor utilizar organismos que estuvieran tan alejados filogenéticamente como sea posible (i.e., un eucarionte, una bacteria, una arquea y un virus). Dicho alejamiento nos permite abarcar caminos evolutivos con divergencia máxima.

El balance entre el azar, la necesidad y las restricciones emergentes también tienen que ver con las elecciones de elaboración del modelo. En este trabajo, por simplicidad,

decidimos utilizar 20 aminoácidos y 64 codones, pues así los códigos generados son comparables, tanto en procesamiento de información como en grado de redundancia, al código actual. En realidad, el número de aminoácidos utilizados por un código es variable. En un trabajo anterior (Mercado, 2009) propusimos un código simplificado y ligeramente ambiguo, el cual utiliza diversos datos, tanto de afinidad estereoquímica como de mecanismos de reacción, y que puede ser interpretado como un código primitivo (Fig. 11). Nuestra elección de formar códigos con las características del código genético estándar, que puede ser observado en nuestros días, responde a la necesidad señalada por Türing (1952) de simplificar y falsificar la realidad en un modelo. Al decidir dejar de lado características fisicoquímicas, afinidades, posibles contingencias históricas y contenido de información funcional, evitamos (o al menos intentamos evitar) hacer lo que puede llamarse "un modelo perfectamente inútil": una representación de la realidad, fiel punto por punto, con un nivel equivalente de complejidad, cuya dificultad de estudio sería idéntica al sistema representado, anulando, por tanto, el objetivo principal de un modelo.

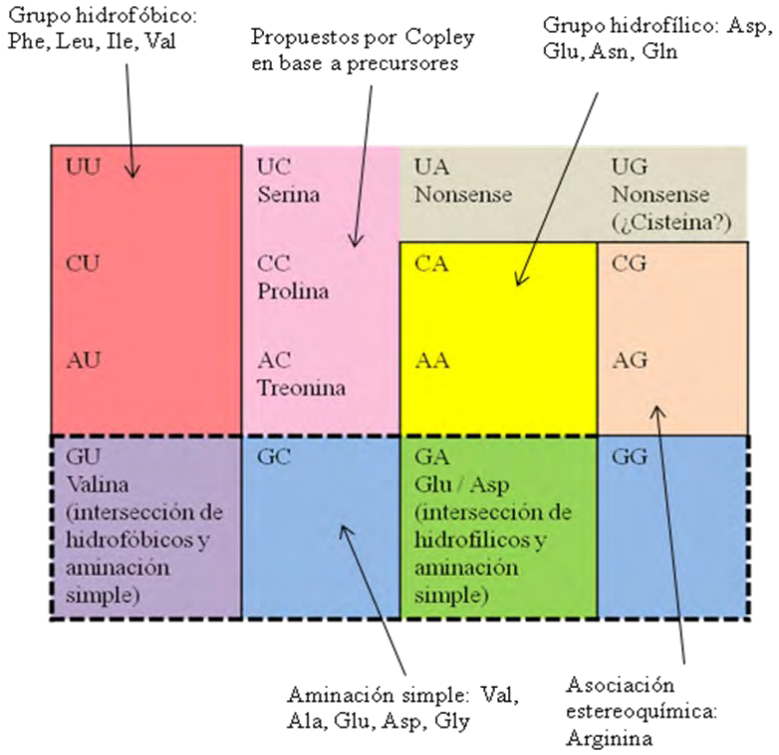

Fig. 11: Propuesta de un código genético primitivo

Propuestas a futuro

El presente trabajo sólo es un acercamiento superficial al código genético y sus propiedades, a la complejidad de la información biológica y a la interacción de los diversos factores que la procesan. Es evidente que son pertinentes investigaciones más profundas sobre algunos de los puntos explorados para afinar el entendimiento acerca de ellos. Algunas preguntas que pueden ser exploradas con más profundidad son las siguientes:

1) *¿La traducción aumenta la complejidad de una secuencia?* En el presente trabajo se ha obtenido mediciones absolutas de compresibilidad mediante un algoritmo determinado, tanto de DNA como de las proteínas codificadas, y posteriormente se les ha comparado para evaluar la compresibilidad relativa; como se ha argumentado a lo largo del trabajo, la compresibilidad se relaciona directamente con la complejidad de una secuencia. Sin embargo, es posible que una secuencia de nucleótidos codificante, únicamente por el hecho de tener que ser sometida al proceso de traducción, inherentemente tenga más complejidad informacional que las secuencias no codificantes. Esta pregunta inicialmente parecería de resolución fácil, pero es necesario considerar que las secuencias no codificantes continúan siendo herméticas en su función y naturaleza. Hablar de DNA no codificante incluye a los intrones, las secuencias de repeticiones en tandem, los fragmentos de regulación, genes no traducibles como rRNA o tRNA, entre muchos otros, y sólo en fechas recientes se ha comenzado a cuestionar la idea de que la mayoría de la información genética que no es traducida es "DNA basura". Para poder acercarse a esta pregunta, es necesario establecer un criterio de funcionalidad que logre trascender la mera traducción, para realizar un análisis equitativo entre genes codificantes y DNA no traducible, y poder comparar la complejidad informacional que ambos tipos de secuencias presentan. Además, en general sólo se puede utilizar genomas eucariontes, pues los procariontes y los virus tienen una tasa muy baja de DNA no codificante.
2) *¿Las relaciones encontradas en este trabajo pueden ser generalizadas?* En el presente trabajo sólo se analizaron cuatro genomas, cada uno muy distinto de los

demás: se escogió un virus de genoma muy grande, la arquea y la eubacteria de genoma más reducido, y un cromosoma de un eucarionte unicelular. Aunque las secuencias utilizadas comprenden miles de pares de bases de ácido nucléico, y se analizaron cientos de genes, el número que alcanzan es sólo una fracción mínima de la diversidad biológica de nuestro planeta. Es necesario extender el estudio a otros organismos y tener una muestra más amplia del espectro informacional de la vida. Si continúa la tendencia observada y se confirman las suposiciones de Abel y Trevors (2005), las secuencias funcionales de los organismos tenderán a tener complejidades muy altas. En tal caso, sería conveniente preguntarse si existe alguna tendencia que determine el nivel de complejidad por grupo filogenético.

3) *¿Cuál es el efecto de las restricciones aumentadas en la generación de códigos y las características informacionales de la traducción?* En la realidad, existe una serie enorme de restricciones fisicoquímicas, informacionales, biológicas e incluso temporales para la estructuración del código genético. Nosotros nos enfocamos únicamente en las restricciones informacionales para elegir los códigos que estudiamos; sin embargo, es altamente probable que la generación de códigos con otro tipo de restricciones arrojen resultados distintos pero complementarios. Dichas restricciones pueden ser, por ejemplo, incluir las características fisicoquímicas o de síntesis de los aminoácidos para aumentar la probabilidad de asignar aminoácidos parecidos en posiciones cercanas; también es posible considerar las reglas de bamboleo (Crick, 1966; Agris et al, 2007) o reglas similares, lo cual implicaría que al asignarse un aminoácido a un codón automáticamente se asignaría a los codones relacionados. Por ejemplo, en los codones NNA, NNU y NNC, la última base se puede aparear con inosina, un nucleótido no estándar, por lo que se podría modelar un sistema en el que los tres codones se asignen simultáneamente. Es posible predecir que los códigos generados con tales reglas tendrán un desempeño similar a los códigos generados aquí, pero los que se acerquen al nivel de complejidad del código genético estándar definitivamente tendrán estructuras muy distintas a las observadas en este trabajo, que analizó códigos generados completamente al azar.

Además de las investigaciones futuras que se han delineado arriba, es posible proponer una serie de líneas de investigación relacionadas con los temas e hipótesis argumentados en esta tesis, pero que no se relacionan directamente con el modelo, por lo que se tendría que especificar una metodología y una serie de argumentos distintos para cada uno de ellas. Por ejemplo:

1) En la primera sección de este texto se bosquejó un posible acercamiento al origen del código genético y, por extensión, probablemente al origen de la vida. Sin embargo, esta propuesta sólo intenta ser eso: un bosquejo. Aunque se expresa las grandes generalidades del escenario, se necesita realizar una investigación de la posible interacción de los tres factores (fisicoquímicos, informacionales y de dinámica poblacional) para proponer un mecanismo más completo de la posible estructuración temprana del código. Esto se tiene que hacer, por supuesto, a la luz de la evolución darwiniana y otras posibles modos de evolución.
2) En este modelo, la parte de dinámica poblacional requiere de un estudio más cuidadoso. Las características fisicoquímicas de los aminoácidos y nucleótidos son bien conocidas y la teoría de Shannon puede ser de gran ayuda para acercarse a la parte informacional, pero hasta el momento hay un número muy reducido de trabajos que intenten relacionar el surgimiento de un lenguaje a través de la interacción de posibles "protolenguajes", es decir, lenguajes y códigos parcialmente estructurados pero que sean flexibles y, por tanto, que puedan llegar a un consenso interpretativo. [12]
3) La escasez de literatura que trata el surgimiento de un código como consecuencia de una interacción "social" hace que existan muchas lagunas por llenar y muchas relaciones por hacer entre temas aparentemente dispares. Por ejemplo, la ya discutida propuesta de Luc Steels (1998, 2007) en la que afirma resolver el problema del establecimiento de símbolos (*symbol grounding problem*, propuesto originalmente en Harnad (1990) mediante robots comunicativos. Debido a que los robots tienen que partir de una base lingüística, como comunicar su acuerdo o

12 El modelo referido en la nota al pie número 10 se basa completamente en una traducción arbitraria. Es necesario, pues, realizar investigaciones más profundas utilizando datos observacionales y no únicamente dinámicas azarosas.

desacuerdo, en realidad no se generan símbolos *de novo*, sino que emergen de una base muy simple. La propuesta de este trabajo posiblemente representa una solución novedosa al problema de Harnad, pues los códigos (y por lo tanto, un sistema lingüístico) se establecen sin convenciones previas. Por supuesto, como lo menciona muchas veces Barbieri (2003, 2008) este tipo de semiótica excluye comportamientos hermenéuticos. Este tema también hace pertinentes nuevas propuestas de mecanismo por los cuales se puede compartir información de un sistema con un código propio a otro; uno de ellos es la propuesta de transmisión horizontal de Woese (Vetsigian et al, 2006), pero no se excluyen escenarios en los que exista una ambigüedad en la traducción de las fases más tempranas.

4) En lo que respecta a la naturaleza de los componentes del código, no existe un mecanismo certero que relacione las propiedades físicas de los componentes de los dos alfabetos. Una posible excepción es el caso de la arginina (Knight y Landweber 2000), uno de los aminoácidos más representados en el código, que además presenta una complementaridad espacial con los codones a los que se asocia. En este sentido, una de las teorías más relevantes es la de Copley et al (2005), que hacen notar la similaridad entre los microambientes fisicoquímicos que causan las diversas partes de los codones y las condiciones necesarias para sintetizar distintos aminoácidos a partir de ácidos tricarboxílicos precursores. Los mecanismos de reacción teóricos que exponen en su artículo relacionan directamente la síntesis específica de aminoácidos, favorecida por los codones que los codifican. Esta propuesta fue el eje central de un trabajo previo (Mercado, 2009). Desafortunadamente, los mecanismos de reacción, aunque completamente plausibles, no han sido llevados a cabo en el laboratorio.

Por último, creemos necesaria la realización de esfuerzos mayúsculos por encontrar las relaciones profundas entre los fenómenos físicos, la información y los seres vivos. Este acercamiento produciría una nueva visión en la manera de acercarse al estudio de la biología de sistemas, e incrementaría nuestro entendimiento del mundo biológico. Como se discute en la sección de “Discusión”, hay puntos en común y conexiones que comienzan a vislumbrarse, como el aumento de complejidad en los procesos informacionales de los

seres vivos, el comportamiento físico de la materia como parte integral (pero no única) de la estructuración del lenguaje biológico y la relación entre la entropía informacional y la termodinámica. Estas líneas de cuestionamiento han sido expresadas por pioneros en los campos de la biosemiótica y de la teoría de la información, como H.H. Pattee (2001), Marcello Barbieri (2003, 2007, 2008) o Hubert Yockey (2005). Es un campo fértil de investigación, como lo muestran los avances de la biosemiótica (Barbieri, 2007) y la perspectiva de considerar a la información como parte de la definición misma de la vida.[13] Sin embargo, es necesario trascender la visión puramente teórica y engarzarla en los procesos biológicos, relacionándola con los posibles modos de origen de la vida, la complejidad metabólica de todos los organismos y los diversos niveles y propiedades emergentes que se encuentran en la riqueza biológica de nuestro planeta.

13 Tal visión es explorada en el manuscrito ya citado de Farnsworth, North, y Gershenson (sin publicar)

Bibliografía citada

Abel, David L. «The capabilities of chaos and complexity». *International Journal of Molecular Sciences* 10.1 (2009): 247-291.

Abel, David L, y Jack T Trevors. «Three subsets of sequence complexity and their relevance to biopolymeric information». *Theoretical Biology & Medical Modelling* 2 (2005): 29.

Agris, Paul F, Franck A P Vendeix, y William D Graham. «tRNA's wobble decoding of the genome: 40 years of modification». *Journal of Molecular Biology* 366.1 (2007): 1-13.

Allison, L et al. «Sequence complexity for biological sequence analysis». *Computers & Chemistry* 24.1 (2000): 43-55.

Ash, Robert. *Information theory*. New York: Dover, 1990.

Barbieri, Marcello. «Biosemiotics: a new understanding of life». *Naturwissenschaften* 95.7 (2008): 577-599.

-----. *Introduction to biosemiotics : the new biological synthesis*. Dordrecht: Springer, 2007.

-----. *The organic codes : an introduction to semantic biology*. Cambridge University Press, 2003.

Battail, Gérard. «Applying Semiotics and Information Theory to Biology: A Critical Comparison». *Biosemiotics* 2.3 (2009): 303-320.

Bedian, Vahe. «Self-description and the origin of the genetic code». *Biosystems* 60.1-3 (2001): 39-47.

-----. «The possible role of assignment catalysts in the origin of the genetic code». *Origins of Life and Evolution of Biospheres* 12.2 (1982): 181-204.

Brooks, Daniel, y EO Wiley. *Evolution as entropy : toward a unified theory of biology*. 2o ed. Chicago: University of Chicago Press, 1988.

Cao, Minh Duc, Trevor I Dix, y Lloyd Allison. «A Simple Statistical Algorithm for Biological Sequence Compression». *2007 Data Compression Conference (DCC'07)*. Snowbird, UT, USA, 2007. 43-52.

-----. «A biological compression model and its applications». *Advances in Experimental Medicine and Biology* 696 (2011): 657-666.

Chaitin, Gregory. «On the simplicity and speed of programs for computing infinite s ets of natural numbers». *Journal of the ACM* 16 (1969): 407-422.

Chandler, Daniel. *Semiotics : the basics*. 2o ed. Routledge, 2007.

Copley, Shelley. D. «A mechanism for the association of amino acids with their codons and the origin of the genetic code». *Proceedings of the National Academy of Sciences* 102.12 (2005): 4442-4447.

Copley, Shelley D., Eric Smith, y Harold J. Morowitz. «The origin of the RNA world: Co-evolution of genes and metabolism». *Bioorganic Chemistry* 35.6 (2007): 430-443.

Crick, Francis «The origin of the genetic code». *Journal of Molecular Biology* 38.3 (1968): 367-379.

-----. «Codon--anticodon pairing: the wobble hypothesis». *Journal of Molecular Biology* 19.2 (1966): 548-555.

Dobzhansky, Theodosius. «Biology, molecular and organismic». *American Zoologist* 4.4 (1964): 443- 452.

-----. «Nothing in biology makes sense except in the light of evolution». *American Biology Teacher* 35 (1973): 125-129.

Durston, Kirk K, David K Y Chiu, et al. «Measuring the functional sequence complexity of proteins». *Theoretical Biology & Medical Modelling* 4 (2007): 47.

Eco, Umberto. *La estructura ausente : introducción a la semiótica*. 1o ed. México D.F.: Debolsillo, 1968.

-----. *Tratado de semiótica general*. 1o ed. México: Debolsillo, 1975.

Emmeche, Claus, y Jesper Hoffmeyer. «From language to nature: The semiotic metaphor in biology». *Semiotica* 84 (1991): 1-42.

Favereau, Donald. «The Evolutionary History of Biosemiotics». *Introduction to biosemiotics: the new biological synthesis*. Springer, 2007. 1-68.

Freeland, Stephen J., y Laurence D. Hurst. «The Genetic Code Is One in a Million». *Journal of Molecular Evolution* 47.3 (1998): 238-248.

Di Giulio, Massimo. «The Coevolution Theory of the Origin of the Genetic Code». *Journal of Molecular Evolution* 48 (1999): 253-254.

-----. «The origin of the genetic code: theories and their relationships, a review». *Biosystems* 80.2 (2005): 175-184.

-----. «The ocean abysses witnessed the origin of the genetic code». *Gene* 346 (2005): 7-12.

Di Giulio, Massimo, y Mario Medugno. «The Historical Factor: The Biosynthetic Relationships Between Amino Acids and Their Physicochemical Properties in the Origin of the Genetic Code». *Journal of Molecular Evolution* 46.6 (1998): 615-621.

-----. «The level and landscape of optimization in the origin of the genetic code». *Journal of Molecular Evolution* 52.4 (2001): 372-382.

Gödel, Kurt. *On formally undecidable propositions of principia mathematica and related systems*. New York: Dover, 1992.

Goodman, Nelson. *Languages of art : an approach to a theory of symbols*. 2o ed. Indianapolis: Hackett, 1976.

Greenwell, Raymond, Nathan Ritchey, y Margaret Lial. *Calculus for the life sciences*. Addison-Wesley, 2003.

Haefner, James. *Modeling biological systems : principles and applications*. New York Chapman & Hall, 1997.

Harnad, Stevan. «The symbol grounding problem». *Physica D: Nonlinear Phenomena* 42.1-3 (1990): 335-346.

Hoffmeyer, Jesper. *Signs of meaning in the universe*. Indiana University Press, 1998.

Hornos, José Eduardo et al. «Symmetry Preservation in the Evolution of the Genetic Code». *IUBMB Life (International Union of Biochemistry and Molecular Biology: Life)* 56 (2004): 125-130.

Hornos, José Eduardo M, Yvone Hornos, y Michael Forger. «Symmetry and symmetry breaking: an algebraic approach to the genetic code». *International Journal of Modern Physics B* 13.23 (1999): 2795-2885.

Jukes, T H, y S Osawa. «The genetic code in mitochondria and chloroplasts». *Experientia* 46.11-12 (1990): 1117-1126.

Kimura, Motoo. *The neutral theory of molecular evolution*. Cambridge Cambridge University Press, 1983.

Knight, RD, y L F Landweber. «Guilt by association: the arginine case revisited». *RNA* 6.4 (2000): 499-510.

Knight, RD, SJ Freeland, y LF Landweber. «Selection, history and chemistry: the three faces of the genetic code.» *Trends Biochem Sci* 24.6 (1999): 241-7.

Knight, Robin D., Stephen J. Freeland, y Laura F. Landweber. «Rewiring the keyboard: evolvability of the genetic code». *Nat Rev Genet* 2.1 (2001): 49-58.

Kolmogorov, AN. «Three approaches to the quantitative definition of information». *Problems of Information Transmission* 1.1 (1965): 1-7.

Kondow, A et al. «An extra tRNAGly(U*CU) found in ascidian mitochondria responsible for decoding non-universal codons AGA/AGG as glycine». *Nucleic Acids Research* 27.12 (1999): 2554-2559.

Kull, Kalevi. «Biosemiotics in the twentieth century: A view from biology». *Semiotica* 127 (1999): 385-414.

Loewenstern, David, y Peter N. Yianilos. «Significantly Lower Entropy Estimates for Natural DNA Sequences». *Journal of Computational Biology* 6.1 (1999): 125-142.

Mercado, José Agustín. *Principios de la información biológica.* Tesis para obtener el grado de biólogo Facultad de Ciencias, UNAM. México, 2009

Monod, Jacques. *Le hasard et la nécessité: essai sur la philosophie naturelle de la biologie moderne*. Paris: Editions du Seuil, 1973.

Morris, Charles. *Foundations of the theory of signs*. University of Chicago Press, 1938.

-----. *Signs, language & behavior*. W W Norton & Co Inc, 1955.

von Neumann, John. *Theory of Self-Reproducing Automata*. University of Illinois Press, 1966.

Nevill-Manning, Craig G, y Ian Witten. «Protein is incompressible». *Data Compression Conference '09* (1999): 257.

Nicolis, Grégoire, y Ilya Prigogine. *Exploring complexity : an introduction*. New York: W.H. Freeman, 1989.

Nöth, Winfried. *Handbook of semiotics*. Bloomington: Indiana University Press, 1995.

Osawa, S et al. «Recent evidence for evolution of the genetic code.» *Microbiology and Molecular Biology Reviews* 56.1 (1992): 229-264.

Pattee, Howard H. «Evolving Self-Reference: Matter, Symbols, And Semantic Closure». *Communication and Cognition - Artificial Intelligence* 12 (1995): 9-27.

---. «The Measurement Problem in Physics, Computation, and Brain Theories». *Nature, Cognition and System II*. Dordrecht: Kluwer, 1992. 179-192.

---. «The physics of symbols: bridging the epistemic cut». *Biosystems* 60.1-3 (2001): 5-21.

Peirce, Charles. *Peirce on signs : writings on semiotic*. Chapel Hill: University of North Carolina Press, 1991.

Prigogine, I. *The end of certainty : time, chaos, and the new laws of nature*. New York: Free Press, 1997.

Rakočević, M. «The genetic code as a Golden mean determined system». *Biosystems* 46 (1998): 283-291.

Rocha, L M. «The physics and evolution of symbols and codes: reflections on the work of Howard Pattee». *Bio Systems* 60.1-3 (2001): 1-4.

Saussure, Ferdinand. *Curso de lingüística general*. Madrid: Akal, 2000.

Schneider, E. «Life as a manifestation of the second law of thermodynamics». *Mathematical and Computer Modelling* 19.6-8 (1994): 25-48.

Sebeok, Thomas A. «Animal communication». *Science (New York, N.Y.)* 147 (1965): 1006-1014.

-----. «Biosemiotics: Its roots, proliferation, and prospects». *Semiotica* 2001.134 (2011): 61-78.

-----. *Contributions to the doctrine of signs*. Lanham MD: University Press of America, 1985.

-----. «The doctrine of signs». *Journal of Social and Biological Systems* 9.4 (1986): 345-352.

Sella, Guy, y David H. Ardell. «The Coevolution of Genes and Genetic Codes: Crick's Frozen Accident Revisited». *Journal of Molecular Evolution* 63.3 (2006): 297-313.

Shannon, Claude. «A mathematical theory of communication». *The Bell System Technical Journal* 27 (1948): 379-423.

Sharov, Alexei A. «Role of Utility and Inference in the Evolution of Functional Information». *Biosemiotics* 2.1 (2009): 101-115.

Sharov, Alexei A. «Functional Information: Towards Synthesis of Biosemiotics and Cybernetics». *Entropy* 12.5 (2010): 1050-1070.

Shcherbak, V. «Arithmetic inside the universal genetic code». *Biosystems* 70 (2003): 187-209.

Smith, E., y Harold J. Morowitz. «Universality in intermediary metabolism». *Proceedings of the National Academy of Sciences* 101.36 (2004): 13168-13173.

Söll, Dieter, y Uttam L. RajBhandary. «The genetic code - Thawing the 'frozen accident'». *Journal of Biosciences* 31.4 (2006): 459-463.

Steels, Luc. «The origins of syntax in visually grounded robotic agents». *Artificial Intelligence* 103.1-2 (1998): 133-156.

-----. «The symbol grounding problem is solved, so what's next?», en *Symbols and embodiment: debates on meaning and cognition.* Oxford University Press, 2008

Stern, Linda et al. «Discovering patterns in Plasmodium falciparum genomic DNA». *Molecular and Biochemical Parasitology* 118.2 (2001): 175-186.

Sugita, T, y T Nakase. «Non-universal usage of the leucine CUG codon and the molecular phylogeny of the genus Candida». *Systematic and Applied Microbiology* 22.1 (1999): 79-86.

Suzuki, C, T Kashiwagi, y K Hirayama. «Alternative CUG codon usage (Ser for Leu) in Pichia farinosa and the effect of a mutated killer gene in Saccharomyces cerevisiae». *Protein Engineering* 15.3 (2002): 251-255.

Suzuki, Takeo et al. «Taurine-containing uridine modifications in tRNA anticodons are required to decipher non-universal genetic codes in ascidian mitochondria». *The Journal of Biological Chemistry* (2011)

Taylor, F. J. R., y D. Coates. «The code within the codons». *Biosystems* 22.3 (1989): 177-187.

Tlusty, Tsvi. «A colorful origin for the genetic code: Information theory, statistical mechanics and the emergence of molecular codes». *Physics of Life Reviews* 7.3 (2010): 362-376.

Toyabe, Shoichi et al. «Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality». *Nature Physics* 6.12 (2010): 988-992.

Türing, Alan. «The Chemical Basis of Morphogenesis». *Phil Trans Roy Soc B* 237.641 (1952): 37-72.

Von Uexküll, Jakob. «The Theory of Meaning». *Semiotica* 42.1 (1982): 25-79.

Vetsigian, K., CR Woese,, y Goldenfeld, N. «Collective evolution and the genetic code». *Proceedings of the National Academy of Sciences* 103.28 (2006): 10696-10701.

Watanabe, Kimitsuna. «Unique features of animal mitochondrial translation systems. The non-universal genetic code, unusual features of the translational apparatus and their relevance to human mitochondrial diseases». *Proceedings of the Japan Academy. Series B, Physical and Biological Sciences* 86.1 (2010): 11-39.

Watson, James, y F H Crick. «A structure for deoxyribose nucleic acid». *Nature* 171 (1953): 737-738.

Winsberg, Eric. «Sanctioning Models: The Epistemology of Simulation». *Science in Context* 12.02 (2008).

Wittgenstein, Ludwig. *Investigaciones filosóficas*. México-Barcelona: Instituto de Investigaciones Filosóficas Universidad Autońoma de México ;Editorial Crítica, 1988.

Woese, Carl R. et al. «Aminoacyl-tRNA Synthetases, the Genetic Code, and the Evolutionary Process». *Microbiology and Molecular Biology Reviews* 64.1 (2000): 202-236.

Wong, J T. «A co-evolution theory of the genetic code». *Proceedings of the National Academy of Sciences of the United States of America* 72.5 (1975): 1909-1912.

Wong, J. Tze-Fei. «Coevolution theory of the genetic code at age thirty». *BioEssays* 27 (2005): 416-425.

Yarus, M. «RNA-ligand chemistry: a testable source for the genetic code». *RNA (New York, N.Y.)* 6.4 (2000): 475-484.

Yarus, Michael, J. Gregory Caporaso, y Rob Knight. «Origins of the genetic code: The Escaped Triplet Theory». *Annual Review of Biochemistry* 74.1 (2005): 179-198.

Yockey, Hubert. *Information theory, evolution, and the origin of life*. New York: Cambridge University Press, 2005.