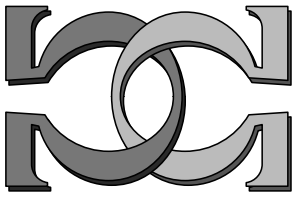
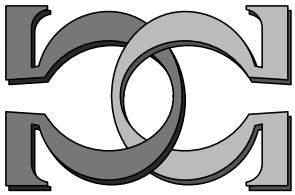
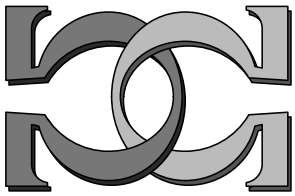
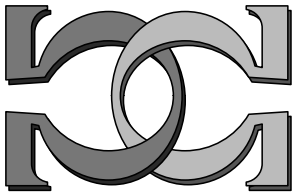


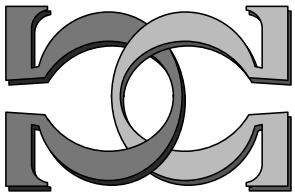




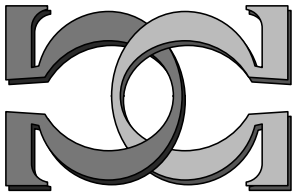
CDMTCS
Research
Report
Series



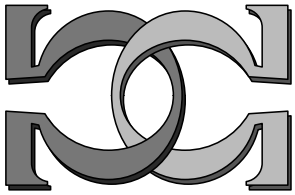
RNA Rings in the Origin of Life



A. Moreira
University of Chile, Santiago, Chile



CDMTCS-224
October 2003



Centre for Discrete Mathematics and
Theoretical Computer Science

RNA Rings in the Origin of Life

Andrés Moreira
Centro de Modelamiento Matemático and Departamento de Ingeniería Matemática
FCFM, U. de Chile, Casilla 170/3-Correo 3, Santiago, Chile
anmoreir@dim.uchile.cl

Abstract

We revisit some ideas proposed earlier by Jacques Demongeot on the origin of life; simulations, combinatorial searches, analysis of tRNA sequences and some additional biological criteria are presented, providing important support for his theory. Starting with some assumptions on the conditions of prebiotic chemistry in primitive Earth, we show –trough simulations– that some particular RNA structures (in particular, some short RNA rings) were more likely to appear and survive. The precise requirements on these sequences lead to an exhaustive combinatorial search of all the possible rings; some additional stability requirements allow the selection of a small number of sequences, and a cluster analysis of them yields a single ring, named the AL ring (for *archetypical loop*), as the most likely winner of the prebiotic game. When the whole list of rings is aligned with a database of tRNAs existing in current organisms, the AL ring turns out to be the one that best matches the free (loop) parts of tRNA, which are believed to be the oldest molecular fossils and are likely to be traces of the first living molecules. Thus we find a possible "missing link" between prebiotic chemistry and the most essential molecules of current organisms. An additional product of the combinatorics is a possible reason for the selection of the start codon and the modern form of genes.

This report corresponds to the third section of the Ph.D. thesis on "Partículas y agentes simples en autómatas celulares y otros sistemas discretos" presented by the author in the Department of Mathematical Engineering of the Universidad de Chile, in August of 2003, under the advice of Dr. Eric Goles and with the collaboration (for this section) of Dr. Demongeot.

Índice general

1. Introducción **8**

2. Nociones básicas de biología molecular **10**

2.1. Componentes básicos de la maquinaria celular 10

2.1.1. Ácidos nucleicos 10

2.1.2. Proteínas 11

2.2. La información genética en la célula 12

3. Teorías sobre los orígenes **18**

3.1. La Tierra primitiva 19

3.2. Escenarios para el origen de la vida 20

3.2.1. Formación de polímeros en el mundo prebiótico 22

3.2.2. Enfoques teóricos 24

3.2.3. La búsqueda de evidencias 24

3.2.4. Fósiles moleculares 25

3.2.5. El tRNA como fósil molecular 26

3.3. Teorías sobre el código genético 28

3.3.1. El accidente congelado 29

3.3.2. Teoría estereoquímica 30

3.3.3. Teorías adaptativas 31

3.3.4. Teorías biosintéticas 31

3.3.5. Código en expansión, código RNY y códigos circulares 32

4. Ideas de J. Demongeot **34**

4.1. Cadena maximal para un código no degenerado 34

4.1.1. Definiciones 34

4.1.2. Cadenas maximales para el código genético 35

4.2. Un RNA minimal en el origen de la vida 36

4.3. Resumen: supuestos, tesis y posibles debilidades 42

5. Simulaciones **44**

5.1. Consideraciones para un código genético simplificado 44

5.2. Simulación espacial 47

5.3. Simulación no espacial 50

5.4. Direcciones a explorar 55

6. Búsqueda combinatoria de RNAs minimales **56**

6.1. Anillos minimales bajo distintas condiciones de codificación 56

6.1.1. A_1: Los 1280 anillos de Demongeot 57

6.1.2. Otros restricciones de codificación 60

6.1.3. A_0: Generalizando los anillos de Demongeot 63

6.2. Hairpins y palindromía 65

6.2.1. Medición de distancias entre anillos 65

6.2.2. Distancias de A_1 y A_0 respecto a sus complementos y AB 69

6.2.3. Formación de hairpins 71

6.2.4. A^*: Selección de 24 anillos 73

6.3. Frecuencias de subcadenas y selección de anillos óptimos 74

7. Relación con secuencias reales **80**

7.1. Los anillos como secuencias primitivas 80

7.1.1. Índice de homogeneidad local 81

7.1.2. GC y redundancia 82

7.1.3. Uso de codones 83

7.2. Datos de tRNAs contemporáneos 84
7.3. Alineamiento de anillos y tRNA 89

8. Síntesis y conclusiones **95**
8.1. Resumen: los anillos bajo nueva luz 95
8.2. Contexto actual de la teoría 100
8.3. Caminos a seguir 101

9. Glosario **104**

10. Bibliografía **107**

11. Anexo 1: Bibliografía adicional sobre el origen de la vida **119**

Índice de cuadros

2.1. Los aminoácidos. 12

2.2. El código genético (casi) universal. 14

5.1. Ejemplo de tabla para simulación no espacial. 50

6.1. Uso de bases en los anillos de Demongeot. 57

6.2. Uso de dinucleótidos en los anillos de Demongeot. 58

6.3. Uso de codones en A_{uaa}. 58

6.4. Ubicación de los aminoácidos en A_{uaa}. 59

6.5. Cantidad de anillos pequeños bajo distintas restricciones de codificación. . 63

6.6. Porcentaje de anillos con las distintas distancias d_H^c respecto de sus complementos. 70

6.7. Porcentaje de anillos con tallos de los distintos tamaños. 72

6.8. Comparación de conjuntos de anillos. 74

6.9. Los 24 anillos de A^*. 75

6.10. Distancia entre los anillos de A^*: d_s y d_h^c. 76

6.11. Distancia entre los anillos de A^*: $\hat{d}_t$. 76

6.12. Promedio de las distancias de los anillos, en A^*. 77

6.13. Ranking de A^* de acuerdo a las diferentes distancias promedio. 77

6.14. Clusters en A^*. 78

7.1. Uso relativo de codones en A_0 y 12 especies. 85

7.2. Cantidad de tRNAs extractados, por grupo y por aminoácido. 87

7.3. Bases conservadas del tRNA. 89

7.4. Esquemas para corte y alineamiento. 90
7.5. Anillos mejor ranqueados al alinear A_0 con la base de datos de tRNA. . . 91
7.6. Alineamiento de AL con *Marchantia* y consenso. 92

Índice de figuras

2.1. Algunas formas de RNA. 11

2.2. Un tRNA típico. 15

2.3. Estructura tridimensional del tRNA. 15

2.4. La fabricación de proteínas. 16

2.5. La duplicación del DNA. 16

3.1. Posibles escenarios en la Tierra primitiva. 19

3.2. Posible desarrollo a través del *RNA world*. 21

3.3. La transferencia horizontal dificulta encontrar las raíces del árbol de la vida. 24

3.4. Posiciones conservadas en el tRNA. 27

4.1. Lectura de la cadena maximal. 35

4.2. Un escenario para el origen de la vida. 37

4.3. Reacciones químicas hipotéticas. 38

4.4. El juego de la vida. 39

4.5. El anillo AB: (a) forma circular; (b) hairpin. 40

4.6. Comparación de AB las partes libres del Gly-tRNA mitocondrial de *Œnothera*. 40

4.7. AB y algunos genomas: gc vs. I. 42

5.1. Un momento en la simulación espacial. 48

5.2. Enlaces de las bases. 48

5.3. Aumento de los anillos a expensas de las cadenas. 50

5.4. Evolución en presencia de un solo aminoácido. 54

5.5. Ventaja de codificar un aminoácido por clase. 54

6.1. Frecuencia de las bases a lo largo de A_1. 59
6.2. Frecuencias de codificación en anillos complementarios. 60
6.3. Frecuencias de codificación en anillos complementarios a A_1 61
6.4. Repeticiones al codificar 20 aminoácidos en largo 22. 61
6.5. Frecuencias de codificación en anillos complementarios a A_0 64
6.6. El mínimo entre dos métricas no es métrica 66
6.7. Histograma de distancia d_H^c respecto a los complementos de los anillos. . . 70
6.8. Histograma de distancia d_H^c respecto al anillo AB. 71
6.9. Partes de un hairpin. 71
6.10. Histograma de tamaño de tallos en hairpins. 73
6.11. Dendrograma para (A^*, d_s), mediante *average clustering*. 79

7.1. IDH para varios conjuntos de anillos. 82
7.2. Ubicación de los anillos y algunos genomas en el plano gc vs. I. 83
7.3. Esquema general del tRNA, en la forma presentada por [Lewin, 2002]. . . 88
7.4. Alineamiento de anillos con tRNAs. 90
7.5. Histograma de pE en A_0 y en un conjunto de secuencias aleatorias. 92
7.6. El hairpin de AL y el alineamiento. 93
7.7. Histograma de alineamientos en A_0. 93

Capítulo 1

Introducción

> Un matemático es un ciego en una habitación a oscuras, buscando un gato negro, que no está ahí.
>
> C. Darwin

Intentando acercarme al estudio de sistemas de partículas reales, empecé un trabajo bajo la guía del profesor Jacques Demongeot, del IMAG, en la Universidad J. Fourier de Grenoble, Francia. Las partículas, en este caso, eran los componentes elementales de la vida: los nucleótidos y los aminoácidos que componen, respectivamente, los ácidos nucleicos y las proteínas. La interacción que nos interesaba no era la que tienen actualmente en los seres vivos, sino más bien la que pudieron tener en el contexto de la sopa primitiva (o pizza, bidimensional) en la que habría comenzado la vida. La forma en que pudo darse el juego de supervivencia de los agregados de estas partículas, y las consecuencias combinatoriales que esto pudo tener sobre las secuencias mismas, fueron el punto de partida. Si bien el carácter de la investigación se alejó rápidamente del estudio formal de las interacciones de estas partículas, derivando más y más hacia el problema biológico, fue la oportunidad para adentrarme en un tema fascinante, como es el origen de la vida.

Muchos investigadores han propuesto ideas en torno al origen de la vida, y del código genético. El tema tiene publicaciones propias, y aparte de la gente que trabaja directamente en él, muchos otros, desde que la ciencia es ciencia, han querido aportar algo con sus ideas. Esta vastedad hace de él un campo lleno de teorías contradictorias, relacionadas en una maraña conceptual y en un cúmulo de información experimental de difícil evaluación. Es en medio de estas agitadas aguas en las que Demongeot situó una idea, un escenario, que ofrece llenar un vacío existente entre las especulaciones sobre la química anterior a la vida, y las formas más primitivas de la maquinaria celular. Se trata, como me hizo notar una amiga bióloga, de algo muy parecido a la paleontología. En esa disciplina, se dispone de fósiles (nunca completos, nunca del todo representativos), y se intenta recrear una historia, nunca segura, pero sí coherente y verosímil, que permita explicar la evolución.

También en la teoría de Jacques Demongeot, revisitada con detalle en esta parte de la Tesis, se tienen fósiles (aunque moleculares), y se busca explicarlos mediante una historia coherente y verosímil, que parte desde la química prebiótica. Incluso aparece, en un rol estelar, una cierta secuencia circular de RNA, el anillo AL (por *Archetypical Loop ring*) al que bien podríamos llamar un "eslabón perdido" en el origen de la vida.

El Capítulo 2 provee el ABC de la biología molecular, imprescindible para que alguien ajeno al área pueda leer lo que viene después. Luego, en el Capítulo 3, se presentan muchas de las ideas que se han propuesto en torno al origen de la vida y del código genético; esto provee el contexto necesario para situar las ideas de Demongeot, y el trabajo realizado en esta Tesis.

La idea de Jacques Demongeot, y su desarrollo, se presenta en el Capítulo 4. Básicamente, plantea un escenario en que cadenas primitivas de RNA habrían sido seleccionadas naturalmente de acuerdo a su estabilidad, que se habría visto favorecida por la afinidad con los aminoácidos circundantes. Basándose en algunas ideas que había planteado anteriormente, sobre la existencia de cadenas generadoras maximales para el código genético, y pensando en los genomas circulares de los organismos más primitivos, propone un genoma minimal, el primero, que habría sido seleccionado en el juego inicial, y del que habrían rastros en las partes libres (básicamente los loops) de los RNA de transferencia de los seres vivos actuales.

Los Capítulos 5, 6 y 7 presentan el trabajo que realicé. El Capítulo 5 presenta un intento por simular el juego primitivo, buscando avalar (o rechazar) las consecuencias que -intuitivamente- Demongeot extrae de él. El Capítulo 6 recoge el estudio de anillos minimales con diversas restricciones de codificación, rigiéndonos por una combinatoria dada por el código genético y sus clases sinónimas de tripletas; se listan conjuntos, se estudian sus estadísticas, se definen métricas para comparar anillos, se agregan algunas ideas biológicas relevantes, etc., hasta llegar a elegir un pequeño conjunto de anillos, y en particular al anillo AL, como candidato más probable a ganador en el juego primitivo. El Capítulo 7, por otro lado, contiene el trabajo realizado con datos reales de secuencias genéticas de organismos actuales. Es ahí donde se "pulen" los fósiles -moleculares-, y se comparan con los anillos, corroborando el rol de eslabón perdido para el anillo AL. Finalmente, en el Capítulo 8 se hace una síntesis de los puntos más importantes del trabajo, y de la forma en que esto ha completado o modificado la tesis del profesor Demongeot; se mencionan también algunas direcciones a seguir a partir de aquí.

Este recuento no agota todo el trabajo que se realizó; hubo mucho tiempo invertido en direcciones que luego se mostraron irrelevantes. Más bien, es el recuento de todo aquello que fue tomando su lugar en el contexto de la teoría, y que de paso me fue convenciendo de que teníamos algo importante entre manos (habiendo partido de un cierto escepticismo). Aún si la explicación que presentamos resulta a la larga errónea, lo que de todos modos se ha demostrado es que hay algo, una coherencia, una serie de "coincidencias", que reclaman una explicación.

Capítulo 2

Nociones básicas de biología molecular

2.1. Componentes básicos de la maquinaria celular

Los dos principales tipos de moléculas que intervienen en la constitución y funcionamiento de todos los seres vivos conocidos son los *ácidos nucleicos* y las *proteínas*. Ambos son polímeros, es decir, cadenas formadas a partir de estructuras más pequeñas. Los ácidos nucleicos están básicamente a cargo del almacenamiento, transmisión y decodificación de información (lo cual incluye la síntesis de las proteínas). Las proteínas, por otra parte, realizan casi todos los trabajos de la célula o el organismo: dan estructura, transmiten mensajes, catalizan reacciones, manejan la energía, etc. Las otras moléculas importantes, de las que no nos ocuparemos, son también polímeros: los polisacáridos (formados por azúcares) y los lípidos (formados por ácidos grasos).

2.1.1. Ácidos nucleicos

Las unidades menores que forman los ácidos nucleicos son los *nucleótidos*, formados por un azúcar, un fosfato y una base. Según el tipo de azúcar los ácidos se clasifican en *ribonucleico* (RNA, que contiene ribosa), y *desoxirribonucleico* (DNA, que contiene desoxirribosa). Las bases son de dos tipos, pirimidinas (timina, uracilo y citosina) y purinas (adenina y guanina). Usualmente se hace referencia a ellas a través de sus iniciales, **t**, **u**, **c**, **a** y **g**. El DNA contiene **a**, **c**, **g** y **t**, mientras que en el RNA la **t** se reemplaza por **u**.

En el caso del DNA, la molécula está formada por dos cadenas, que se enlazan por *hibridación*, a través de un pareo base a base que da origen a la ya clásica doble hélice de Watson y Crick. Por razones energéticas, sólo son estables los pareos **g-c** y **t-a** (o **a-u**, en el caso de pareos de RNA; ocasionalmente se da **g-u**), de modo que las dos hebras de una doble hélice presentan una simetría dada por estos pares. Existen también moléculas

de DNA formadas por una sola hebra, que a veces se cierra en un círculo.

El RNA, en cambio, suele permanecer como una única cadena abierta, cerrándose a veces circularmente. Sin embargo, los pareos de Watson y Crick también son importantes, pues a través de ellos, plegándose sobre sí mismo, el RNA toma ciertas formas particulares que son importantes para sus funciones, para su estabilidad, y también para su proceso de edición (algunas posibles formas se ven en la Figura 2.1). A estos plegamientos se les llama *estructura secundaria*; el término hace referencia a la dimensión del ordenamiento. Así, la estructura *primaria* se refiere al arreglo lineal de bases (la secuencia de "letras"), y la estructura *terciaria* al arreglo tridimensional.

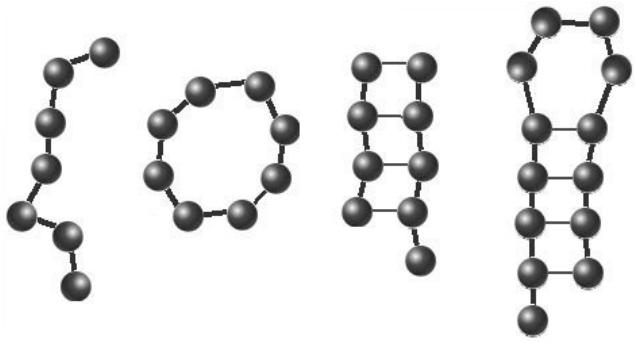


Figura 2.1: Algunas formas de RNA.

Este último nivel de estructura es también importante, pues las distintas configuraciones tienen distintas funcionalidades y estabilidades. Al estar las estructuras 3D y 2D determinadas por la 1D, las restricciones que la evolución imponga sobre las configuraciones de dimensión mayor influyen en la conformación de las secuencias lineales.

2.1.2. Proteínas

Las proteínas, por otro lado, están compuestas por monómeros llamados *aminoácidos*, de los cuales existen 20 variedades en los seres vivos. Los aminoácidos se enlazan mediante *enlaces peptídicos*; se llama péptidos a las cadenas pequeñas de aminoácidos. También en este caso la estructura tridimensional es importante, y lo es más aún que en el caso de los ácidos nucleicos; con frecuencia la forma determina por completo la función de la proteína.

El Cuadro 2.1 muestra las abreviaturas estándar para los aminoácidos, así como también las letras que se les asocian. La clasificación que se presenta es la más usual, y describe la naturaleza de la cadena lateral de cada aminoácido (que determina su carga iónica y por lo tanto su contribución a la estructura de la proteína).

Tipo	Nombre	Abreviatura	Codificación
Neutros e hidrófobos	Alanina	Ala	A
	Valina	Val	V
	Leucina	Leu	L
	Isoleucina	Ile	I
	Prolina	Pro	P
	Triptofano	Trp	W
	Metionina	Met	M
	Fenilalanina	Phe	F
Neutros y polares (hidrófilos)	Cisteína	Cys	C
	Glicina	Gly	G
	Tirosina	Tyr	Y
	Serina	Ser	S
	Treonina	Thr	T
	Asparagina	Asn	N
	Glutamina	Gln	Q
Básicos	Lisina	Lys	K
	Arginina	Arg	R
	Histidina	His	H
Ácidos	Ácido glutamínico	Glu	E
	Ácido aspártico	Asp	D

Cuadro 2.1: Los aminoácidos.

2.2. La información genética en la célula

Todos los organismos vivos conocidos, así como los virus, viroides y fagos, codifican su información mediante ácidos nucleicos. En el caso de los viroides, el organismo no es más que RNA desnudo; en el caso de un virus, una molécula de RNA o de DNA forma el organismo en conjunto con una proteína. En todos los demás casos (bacterias, animales, vegetales) el organismo está formado por células. No nos detendremos en el caso de los virus y viroides, sino en el de los organismos celulares.

Existen dos tipos de célula: la *procariota*, que no posee un núcleo definido, y la *eucariota*, que sí lo posee. Esta división es el primer principio de clasificación de los seres vivos, pues está ligada a una serie de otras diferencias (incluidas algunas en la estructura y composición de la información genética). Procariotas son básicamente las bacterias, divididas en dos grandes "reinos" (las arqueobacterias, *Archaea*, y las eubacterias), mientras que protistas, hongos, animales y plantas son eucariotas (*Eukarya*). El DNA se encuentra siempre en el núcleo, en los *cromosomas*; en general un cromosoma es una única molécula de DNA, asociada a un armazón de proteínas. El RNA se presenta en varios tipos distintos, dentro y fuera del núcleo.

Se llama *gen* a la unidad básica de información genética; a *grosso modo*, los genes son los segmentos de DNA que codifican los distintos RNA, y en particular, los que codifican al RNA mensajero y por lo tanto a las distintas proteínas.

El proceso de síntesis de las proteínas comienza en el núcleo, donde algunas moléculas gatillan el movimiento de *transcripción*, a través del cual se forma una cadena de RNA *mensajero* (mRNA), que copia un segmento de una hebra de DNA (poniendo **c** frente a **g**, **g** frente a **c**, **u** frente a **a**, **a** frente a **t**). El lugar de inicio de la transcripción está indicado por ciertas secuencias especiales (*promoters*; para detalles, puede consultarse [Lewin, 2002] o [Tracqui y Demongeot, 2003]). En general es una sola hebra la que codifica, pero con frecuencia ambas lo hacen, incluso -raramente- en forma simultánea (es decir, un "texto" es usado para codificar una proteína, y parte de su "negativo" es usado en sentido contrario para codificar otra). También se da el caso de traslape entre dos secuencias codificadoras de una misma hebra, sobre todo en procariotas y en virus, cuyo genoma ha evolucionado para volverse más compacto.

El mRNA es muchas veces "editado", en un proceso que se conoce como *splicing*: a través de enzimas (y a veces también en ausencia de ellas) algunos trozos son extraídos de la secuencia, hasta que queda la secuencia que realmente codifica la proteína. Se llama *intrones* a los segmentos que se descartan, y *exones* a los demás. Pese a que los intrones tienen ciertas marcas características, no siempre es posible distinguirlos de manera unívoca en las secuencias: incluso puede darse el caso de que una misma secuencia codifique proteínas distintas sólo gracias a un splicing distinto (la presencia de alguna enzima en cierta concentración puede actuar como un interruptor que determina lo que se produce). Si bien el fenómeno del splicing es ubicuo en los eucariotas, se pensaba que este splicing alternativo era una rareza; tras el reciente completamiento del genoma humano, se ha observado que también es algo muy frecuente (se da en alrededor del 35 % de los genes, y de hecho, contribuye en gran medida a la plasticidad del sistema inmune de los mamíferos [Pasqual et al., 2002]).

Existe una vieja controversia respecto a la antigüedad de los intrones: para algunos, son una innovación que apareció a lo largo de la historia de la vida; para otros, estuvieron presentes desde el principio, y su notoria rareza en los procariotas se debe a que éstos los han perdido al evolucionar hacia genomas más compactos.

El mRNA ya editado sale del núcleo y se dirige a los *ribosomas*, estructuras a su vez formadas por RNA (del tipo llamado *ribosomal*, rRNA) y proteínas, para su *traducción*. Los ribosomas son "fábricas" de proteínas: aquí el mRNA se va haciendo correr, de la manera en que se lee una cinta, y se va formando una proteína cuyos aminoácidos obedecen a la secuencia indicada por el mRNA, siguiendo lo que se llama el *código genético* (prácticamente idéntico en todos los organismos), detallado en el Cuadro 2.2.

En este código, los aminoácidos son codificados por tripletas de bases ("codones"). Naturalmente existe redundancia (es decir, el código es "degenerado"), por lo cual algunos aminoácidos son codificados por más de un codón. La redundancia suele estar en la tercera posición del codón (en casi todos los casos, el sentido no cambia al hacerse una transición

codón	aá	codón	aá	codón	aá	codón	aá
aaa	*K*	**caa**	*Q*	**gaa**	*E*	**taa**	STOP
aac	*N*	**cac**	*H*	**gac**	*D*	**tac**	*Y*
aag	*K*	**cag**	*Q*	**gag**	*E*	**tag**	STOP
aat	*N*	**cat**	*H*	**gat**	*D*	**tat**	*Y*
aca	*T*	**cca**	*P*	**gca**	*A*	**tca**	*S*
acc	*T*	**ccc**	*P*	**gcc**	*A*	**tcc**	*S*
acg	*T*	**ccg**	*P*	**gcg**	*A*	**tcg**	*S*
act	*T*	**cct**	*P*	**gct**	*A*	**tct**	*S*
aga	*R*	**cga**	*R*	**gga**	*G*	**tga**	STOP
agc	*S*	**cgc**	*R*	**ggc**	*G*	**tgc**	*C*
agg	*R*	**cgg**	*R*	**ggg**	*G*	**tgg**	*W*
agt	*S*	**cgt**	*R*	**ggt**	*G*	**tgt**	*C*
ata	*I*	**cta**	*L*	**gta**	*V*	**tta**	*L*
atc	*I*	**ctc**	*L*	**gtc**	*V*	**ttc**	*F*
atg	*M*	**ctg**	*L*	**gtg**	*V*	**ttg**	*L*
att	*I*	**ctt**	*L*	**gtt**	*V*	**ttt**	*F*

Cuadro 2.2: El código genético (casi) universal.

en esta posición –reemplaza de la base por otra del mismo tipo, purina o pirimidina– o incluso al hacerse una transversión –reemplazo por una del otro tipo); la posición más importante parece ser la segunda, al parecer correlacionada con la naturaleza hidrofóbica o hidrofílica del aminoácido codificado.

Tres codones indican el término de una cadena (**taa**, **tag**, **tga**), mientras que uno (**atg**) está al comienzo de toda secuencia codificadora (sin embargo, como codifica a la metionina (*M*), también puede darse al interior de la secuencia, por lo que su presencia no indica necesariamente un punto de inicio). La forma de un gen típico se muestra a continuación:

aug	**ccu**	**auu**	**aug**	...	...	**ugu**	**gca**	**acg**	**uga**
M	*P*	*S*	*M*	**...**	...	*C*	*A*	*T*	stop

Los aminoácidos son traídos por un tercer tipo de RNA, el de *transferencia* (tRNA), que tiene una forma de trébol muy característica, y en el que algunas bases son modificadas (y transformadas en bases "atípicas") después de su formación. Un tRNA típico puede verse en la Figura 2.2; $3'$ y $5'$ indican la dirección de los RNA (determinada por la ubicación de sus bases en relación al esqueleto de fosfatos). El tRNA se enlaza por un lado a un aminoácido y por el otro a un codón correspondiente, de modo que al alinearse los tRNA con el mRNA se va formando, por detrás de los tRNA, la cadena de la proteína. A esta secuencia, bosquejada en la Figura 2.4, se le suele llamar el "dogma fundamental de la biología molecular":

$$\text{DNA} \longrightarrow \text{RNA} \longrightarrow \text{PROTEÍNA.}$$

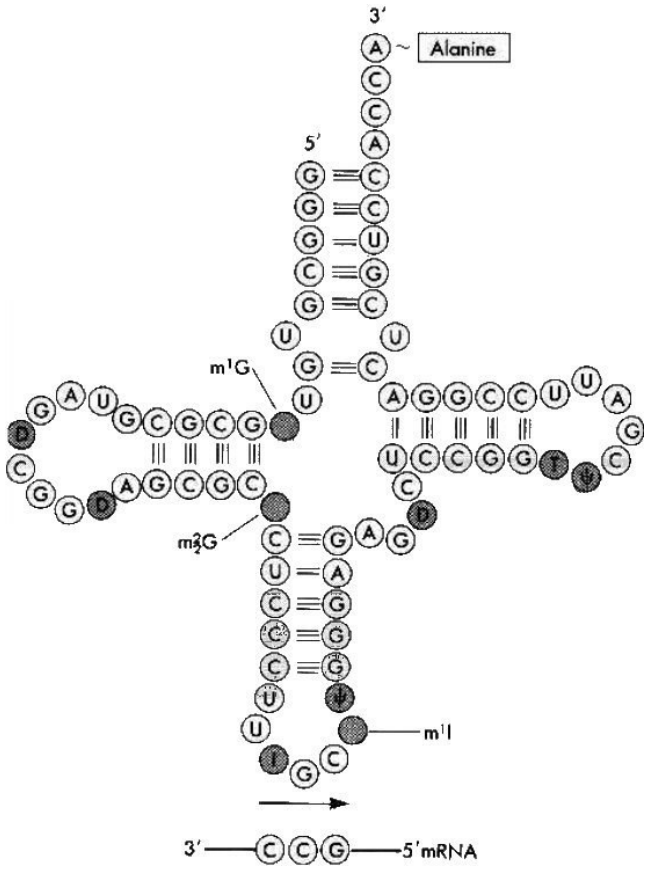


Figura 2.2: Un tRNA típico.

Figura 2.3: Estructura tridimensional del tRNA.

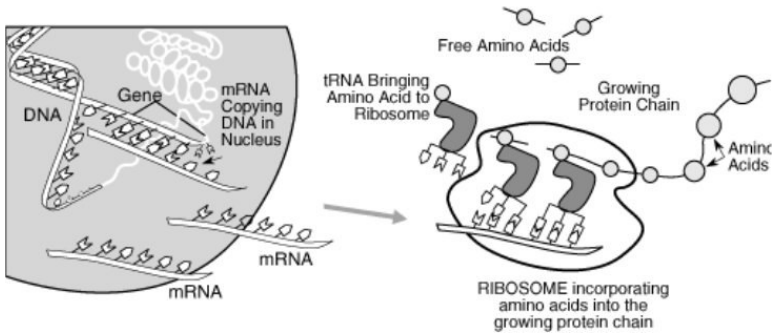


Figura 2.4: La fabricación de proteínas: requiere DNA, RNA y otras proteínas.

Este esquema sólo duró como "dogma" durante una década, hasta el descubrimiento de la *retrotranscriptasa*, capaz de revertir el primero de los pasos; esta enzima puede ser muy antigua, pues es codificada por los intrones del grupo II. Por otro lado, en [Nashimoto, 2001] se propone un mecanismo a través del cual pudo haberse revertido incluso el segundo paso. Sin embargo, el "dogma" subsiste como la representación de la abrumadora mayoría de los casos.

Las moléculas de tRNA y rRNA son también codificadas por el DNA; la función codificadora del DNA consiste exclusivamente en la codificación de mRNA, rRNA, tRNA, y algunos otros tipos de RNA más pequeños. El procedimiento para la síntesis de todos estos RNA es similar al del mRNA. En lo que respecta al DNA, es generado mediante la *duplicación* de un DNA preexistente, como muestra la Figura 2.5. Se trata de un proceso bastante complejo, que requiere de una serie de enzimas (proteínas) para ser llevado a cabo.

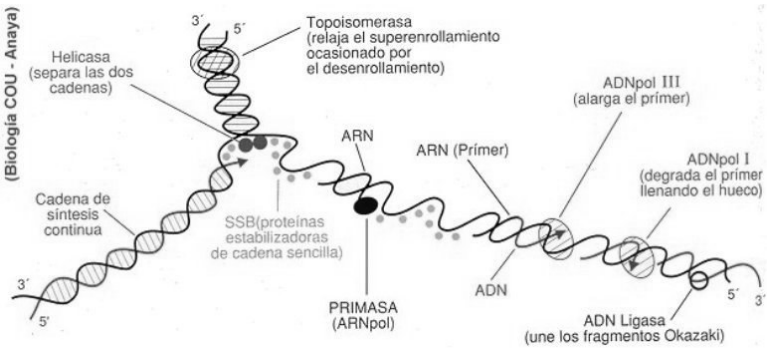


Figura 2.5: La duplicación del DNA: requiere de diversas proteínas.

Es importante mencionar que el DNA nuclear no es el único que está presente en las células eucariotas: en los organelos (mitocondrias y cloroplastos), estructuras que funcionan prácticamente como "sub-células", existe un DNA propio, procesado por una maquinaria idéntica a la ya descrita; en algunos casos los genomas del organelo y de la célula ("DNA

citoplasmático") se interrelacionan: hay genes que se usan en el citoplasma y son codificados en una mitocondria, y viceversa. Se piensa que se trata de bacterias que en algún momento remoto entraron en simbiosis con la célula eucariota primitiva, convirtiéndose finalmente en parte de ella. Es en las mitocondrias donde se presentan las pocas variaciones conocidas respecto del código genético universal.

Capítulo 3

Teorías sobre los orígenes

El origen de la vida es un tema fascinante, y ha sido objeto de debate desde que se tiene uso de razón (aún en los tiempos en que era remitido a un demiurgo). Es por esto que muchos investigadores lo han tomado como centro de su trabajo, y muchísimos más han buscado aportar ideas al respecto. El resultado es una literatura gigantesca, llena de ideas que van desde lo muy razonable hasta lo disparatado; como se dice en [Maher, 2002], "pónganse 10 teóricos [sobre este tema] en una sala, y emergerán 12 opiniones". Tal vez la principal sorpresa que se lleva un matemático al entrar a informarse en temas biológicos, y en especial en uno tan llamativo y polémico como éste, es encontrar con frecuencia artículos, cada uno de apariencia sólida y estadísticamente respaldado, que se contradicen flagrantemente; rara vez se dan situaciones semejantes en la literatura matemática, gracias a los distintos criterios usados para validar una idea.

En las páginas que siguen he intentado poner orden a las ideas más importantes que he logrado reconocer en este ámbito, pero por ser yo un forastero, es probable que peque reiteradamente de omisión o de valoración errónea. En algunos casos es conciente: sé que no he hecho justicia a algunos autores importantes (empezando quizás por E. Szathmáry, de quien por fortuna se puede hallar en español un excelente libro, [Maynard Smith y Szathmáry, 2001]), que probablemente he ignorado a algunos que pueden resultar ser seminales, y que he citado a algunos a propósito de un tema cuando probablemente tienen mucho que decir respecto a otro.

Con todo, creo que lo que sigue cubre bastante bien los principales problemas e intentos de respuesta, con el fin de poner en el necesario contexto la propuesta de J. Demongeot, en torno a la que he trabajado. Para una teoría completa, hará falta armar un puzzle que tiene muchas piezas. Creo que la hipótesis de Demongeot es una, y las demás puede que estén en alguna parte de este capítulo.

Como última advertencia quiero mencionar algo leído en la web de *Nature*, a propósito del origen de la vida: puede ser imprescindible ignorar parte de los datos. Un científico citado allí señala que "para crear una hipótesis coherente, tenemos que decir 'estos datos no calzan, pero seguiremos adelante de todos modos′ ".

3.1. La Tierra primitiva

Hay amplias controversias sobre el escenario en que la vida debió originarse. Se sabe que debió ocurrir hace menos de 3.800 millones de años (pues recién entonces la temperatura bajó a un rango -habitable- de entre 50 y 90 grados), y hace más de 3.600 millones de años (edad de los indicios más antiguos de vida); esto indica un inicio relativamente rápido, tan pronto como las condiciones estuvieron dadas (aunque existe debate respecto a estas fechas). Descartando la intervención divina y la panspermia (arribo de vida desde fuera del planeta), la vida debería haber surgido en un mundo caliente, violento, y con una atmósfera compuesta principalmente por dióxido de carbono y nitrógeno, con una importante presencia de agua. También debió haber hidrógeno, así como algunos compuestos: H_2S, NH_3 (amoníaco) y CH_4 (metano); sin embargo, el tema es materia de debate (por ejemplo, en [Lazcano, 1997; Nisbet y Sleep, 2001]), y la visión predominante varía cada lustro. Se trata de una atmósfera reductora, en la que la formación espontánea de moléculas ricas en carbono no habría costado tanta energía como la que costaría en la actual. Dentro del planeta se han sugerido diversos lugares como posible cuna de la vida: en las profundidades de la tierra, en arcilla, en fumarolas submarinas, y sobre todo, en las tierras situadas al borde del océano (ver Figura 3.1); ya Darwin, premonitoriamente, habló de *a warm little pond* ("una pequeña laguna tibia"), que tal vez cambiaríamos hoy por *a warm little beach*.

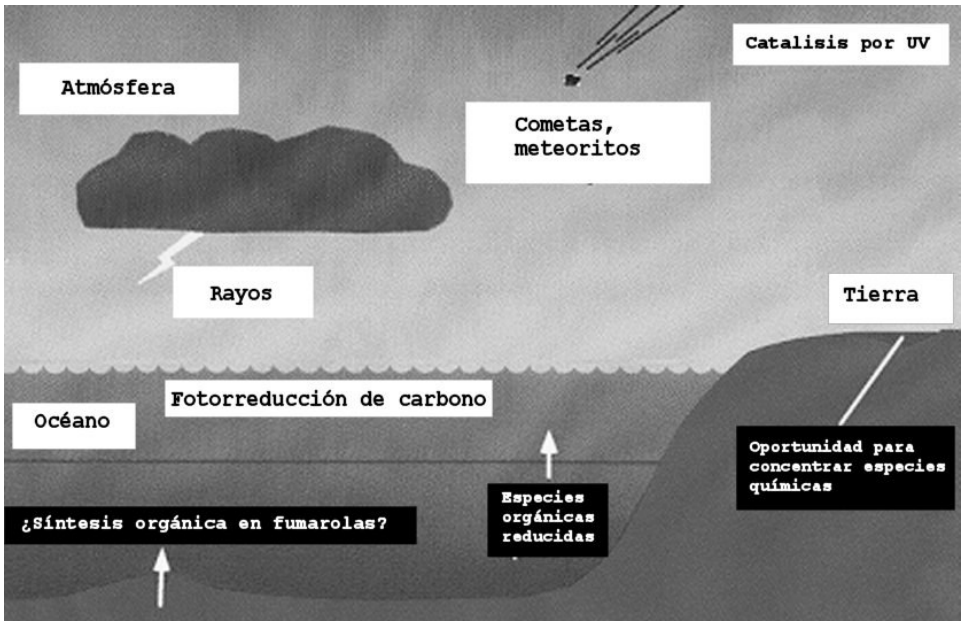


Figura 3.1: Posibles escenarios en la Tierra primitiva.

En un clásico experimento de 1953, S. Miller y H. Urey mostraron que una atmósfera como la que se supone existió, y sometida a condiciones verosímiles, puede generar espontáneamente las moléculas esenciales para la vida; en esta experiencia y en otras posteriores se

pudieron formar todos los aminoácidos y todas las bases (purinas y pirimidinas). Sobre la abundancia relativa de estos productos existe, otra vez, considerable debate. Se concuerda sí en que la síntesis de nucleótidos completos debió ser más difícil (aunque posible); en general, la formación de polímeros no parece fácil, y por esto se ha sugerido que las reacciones tuvieron lugar en algún medio físico que las protegiera. En experimentos con superficies de arcilla se han logrado polinucleótidos y polipéptidos de hasta 50 unidades de longitud (y se habla de una "pizza" más que de una "sopa" primitiva, pues muy probablemente las reacciones se desarrollaron en un escenario bidimensional) . Otros han sugerido un inicio sin polímeros (ver más abajo).

3.2. Escenarios para el origen de la vida

Se han propuesto una multitud de posible orígenes para la vida, tanto desde el punto de vista de la factibilidad de las reacciones químicas, como desde el punto de vista teórico (las condiciones necesarias para la emergencia de un sistema con las características que conocemos de la vida).

Algunas de estas teorías prescinden de polímeros. Por ejemplo, S. Sowerby [Sowerby y Heckl, 1998; Sowerby et al., 2002], basándose en trabajo experimental, sugiere que las purinas y pirimidinas pudieron haberse organizado espontáneamente en capas bidimensionales, a las que se habrían ligado los aminoácidos. D. Segré [Segré et al., 2000; Segré et al., 2001] ha descrito un escenario de *compositional genomes* en que "bolas" compuestas de lípidos interactúan, se reproducen, etc. G. Wächtershäuser [Wächtershäuser, 1994; Wächtershäuser, 2000] ha propuesto que el origen de la vida habría sido mediante reproducción "análoga" (en contraposición a la "digital" de los polímeros), de cadenas de reacciones metabólicas.

Una idea un poco minoritaria, pero ciertamente interesante, es la de un origen no orgánico. La tesis fue propuesta a fines de los 60 por A. Cairns-Smith [Cairns-Smith, 1982], que sostiene que en un medio formado por montmorillonita (una arcilla que se da en la región de Montmorillon, Francia) pudo haber replicación de "imperfecciones" o arreglos electrónicos, en capas sucesivas, y que estas estructuras pudieron haber catalizado las siguientes reacciones entre moléculas orgánicas para inventar el metabolismo y los polímeros que habrían "invadido" luego el sistema.

Sin embargo, la mayor parte de las propuestas postulan un escenario centrado en polímeros, ya sean proteínas, ácidos nucleicos, o ambos. Hay un debate importante respecto al orden en que estos debieron aparecer: se trata de una nueva versión del clásico problema del huevo y la gallina, pues como vimos antes, hoy en día las proteínas son esenciales para la reproducción de los ácidos nucleicos, y estos últimos lo son para la síntesis protéica.

Mundo de RNA: La visión predominante a la fecha habla de una primera etapa en la que el RNA habría sido la "forma de vida" principal; la necesidad de una molécula con

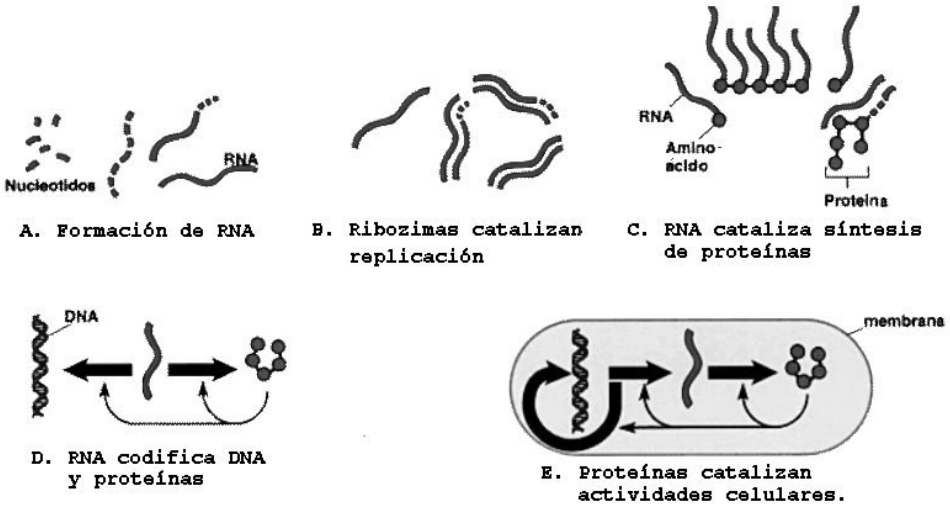


Figura 3.2: Posible desarrollo a través del *RNA world*.

información hereditaria se esgrime como primera razón para situar al RNA en el origen. Un argumento importante a favor de esta tesis fue el descubrimiento por T. Cech en 1982 de las ribozimas, moléculas de RNA con capacidad catalítica. No sólo se han sintetizado ribozimas en laboratorios, sino que se las ha encontrado en diversos rincones de la maquinaria celular de los seres vivos, incluyendo estructuras que se cree muy antiguas (un recuento de ribozimas naturales se encuentra en [Doudna y Cech, 2002] y una discusión de ribozimas artificiales en [Jäschke, 2001]). Con esto, es posible imaginar un mundo en que los roles que actualmente cumplen las proteínas eran llevados a cabo por ácidos nucleicos, y también es posible pensar en rutas evolutivas a través de las cuales las proteínas habrían ido reemplazando progresivamente al RNA debido a sus mayores capacidades [Garcia y Weeks, 2003; Ikawa et al., 2003] (de hecho, subsisten formas intermedias, las ribonucleoproteínas -RNP- formadas por RNA **y** proteína, entre las que se destaca la ribonucleasa P, que interviene en la formación del tRNA). El hecho de que el corazón de la maquinaria vital esté en manos de RNAs también avala esta hipótesis; no hay razón *a priori* para que, por ejemplo, la función del tRNA sea cumplida por un RNA, y como señaló una vez Francis Crick al ver su compleja estructura tridimensional, "parece el intento de la naturaleza para hacer el trabajo de una proteína mediante un ácido nucleico" (véase la comparación entre un tRNA y una proteína en [Landweber, 1999], que también incluye una discusión y catálogo de ribozimas). Ya en los 60 se especuló sobre este tema [Crick, 1968], pero sólo después del descubrimiento de las ribozimas empezó a recibir creciente atención. El término mismo de "RNA world" fue acuñado por [Gilbert, 1986].

Mundo de proteínas: Los críticos del *RNA world* destacan la mayor capacidad catalítica de las proteínas, y la mayor facilidad con que emergen espontáneamente en el caldo primigenio, en contraste con los complejos nucleótidos. Aunque es más difícil imaginar formas de reproducción de proteínas que de RNA (ver sección siguiente), en ausencia

de un sistema de pareos del tipo de Watson-Crick, entre los proponentes de este modelo se cuentan voces tan destacadas como la de M. Eigen [Eigen, 1992]. Una versión reciente, en el marco de una ambiciosa teoría, es presentada en [Ikehara, 2002b; Ikehara, 2002a].

Mundo de PNA: Por ser el RNA relativamente complejo e inestable, esta teoría prefiere asumir que la primera etapa habría sido la del ácido peptídico-nucleico (PNA), más estable y también capaz de formación espontánea y de autorreplicación [Nielsen, 1993; Böhler et al., 1995; Knight y Landweber, 2000]. Se han propuesto también otros precursores [Schwartz, 1997].

Hipótesis mixtas: Algunos aducen la necesidad de una coevolución de proteínas y ácidos nucleicos desde el principio; en ese escenario, las mejores virtudes de cada uno se habrían aprovechado en pequeños colectivos moleculares con capacidad autorreproductora. [Dyson, 1985], por ejemplo, propone un doble origen de la vida (el metabolismo en la proteínas, y la reproducción en los ácidos nucleicos), y [Nashimoto, 2001] da evidencia experimental para la factibilidad de un escenario de coevolución en que tRNAs y tRNAs "inversos" traducen entre un mundo y el otro.

Otros elementos de discusión son el momento y la forma de aparición del DNA, así como el origen de la estructura celular: para muchos, la existencia de algún tipo de pared o membrana es esencial para que se produzca evolución, pues se requiere protección respecto al medio y compartimentación que proporcione "identidad". Se han propuesto algunas formas en que estas membranas pudieron generarse (mediante lípidos o proteínas) y también se ha sugerido que las reacciones pudieron ocurrir en medios físicos ya compartimentados.

3.2.1. Formación de polímeros en el mundo prebiótico

Muchos experimentos en los últimos años han apuntado a determinar las posibles formas de reproducción de biopolímeros en el mundo prebiótico. En la bioquímica actual, la polimerización requiere de enzimas, tanto en el caso de los ácidos nucleicos como en el de las proteínas. Para el mundo primitivo, se busca demostrar la posibilidad de actividad enzimática por parte de polímeros pequeños, o de replicación por plantillas ("templates"), es decir, a través del pareo entre un polímero y monómeros libres (o polímeros más cortos). Hasta ahora los mayores avances han sido con combinaciones de ambos efectos.

Síntesis de péptidos

Mediante evolución *in vitro* se han hallado ribozimas capaces de catalizar la formación de algunos péptidos [Zhang y Cech, 1997]. Por otro lado, también es posible que los propios péptidos hicieran este trabajo, considerando su mayor capacidad enzimática. El grupo de D. Lee, K. Severin, M. Ghadiri y otros colaboradores ha explorado sistemas

auto-replicativos de péptidos, con plantillas y actividad enzimáticas, usando un enfoque de hiperciclos [Lee et al., 1996; Lee et al., 1997b; Lee et al., 1997a; Severin et al., 1997a; Severin et al., 1997b] (véase también un comentario [Kauffman, 1996] al artículo de Lee [Lee et al., 1996], con título "Even peptides do it"). El sustrato físico puede haber sido importante también como catalizador [Bujdák y Rode, 1996a; Bujdák y Rode, 1996b; Kochavi et al., 1997]. En algunas bacterias actuales (e.g., *Bacillus brevis*) se han encontrado proteínas sintetizadas por otras proteínas, aunque nada indica que el mecanismo sea antiguo.

En general se tiende a descartar la reproducción por efecto únicamente de plantillas en el caso de los péptidos (siendo éste uno de los argumentos contra el "mundo de proteínas") pero hay indicios de que también puede ser posible en algún grado. L. Harris [Harris et al., 1999] habla de esta posibilidad, citando el trabajo de J. Blalock [Blalock y Smith, 1984; Blalock y Bost, 1986; Blalock, 1990; Clarke y Blalock, 1991], quien ha estudiado la relación entre algunas proteínas y sus "antiproteínas", proteínas codificadas por la hebra complementaria de DNA: parece haber afinidad entre ambas. Este tema tiene su origen en [Mekler, 1969], y algunos reviews recientes son [Tropsha et al., 1992], [Jarpe y Blalock, 1994], [Forsdyke, 1995] y [Root-Bernstein y Holsworth, 1998]; parte del interés se debe a posibles aplicaciones médicas, utilizando la complementaridad de péptido con antipéptido para bloquear reacciones. Sin embargo, sigue habiendo debate sobre la existencia de estas afinidades, y en caso de existir, tampoco es claro que puedan contribuir a un pareo que conduzca a la reproducción; algunos, como [Houen, 1999], lo descartan de plano. Una propuesta de autorreplicación de proteínas, con características distintas a los autores antes citados, aparece en las teorías de K. Ikehara [Ikehara, 2002b; Ikehara, 2002a].

Síntesis de RNA

Debido a los pareos de Watson y Crick, es relativamente fácil imaginar una reproducción por plantillas en el caso de los ácidos nucleicos, sobre todo sobre sustratos de arcilla que pudieron facilitar la formación de los puentes de hidrógeno de la hibridación. Algunos estudios de replicación por plantillas prescindiendo de actividad enzimática aparecen en [Gao y Orgel, 2000], [Mellersh y Wilkinson, 2000] y [Assouline et al., 2001].

En cuanto a ribozimas capaces de catalizar la polimerización, varias se han encontrado, por lo general combinando las ribozimas con plantillas, como se puede ver en [Johnston et al., 2001], en el trabajo del grupo de J. Szostak [Szostak, 2003] y también en el grupo de G. Joyce (por ejemplo, [Wright y Joyce, 1997], [McGinness y Joyce, 2002] y [Paul y Joyce, 2002], que incluye una ribozima auto-replicadora; véanse también los dos reviews [Joyce, 2002] y [McGinness y Joyce, 2003]).

3.2.2. Enfoques teóricos

En cuanto a la forma en que debieron darse las reacciones, también se han propuesto diversas ideas. Algunos postulan que una molécula simple de RNA pudo tener la capacidad para autorreplicarse; para otros, debió tratarse de un conjunto de moléculas, unas contribuyendo a la formación de otras (un conjunto *autocatalítico*). Según S. Kauffman [Kauffman, 1993], toda química suficiente compleja producirá tarde o temprano sistemas de este tipo, dando origen a la vida. El concepto minimal de vida llamado *autopoiesis*, acuñado por Maturana y Varela [Varela et al., 1974; Varela, 2000], incluye la idea de autocatálisis sumada a una frontera (membrana), exigiendo que ésta también sea generada por las reacciones (el sistema produce todos sus componentes); ha sido aplicado al tema del origen de la vida, por ejemplo, en [Luisi, 1993].

M. Eigen [Eigen y Schuster, 1978; Eigen et al., 1981] introdujo dos nociones importantes: en primer lugar, la de *cuasiespecies*, que describe comunidades de secuencias, no idénticas sino similares, producto de replicación con errores. Tras mostrar que la baja fidelidad de la copia limita el tamaño de las secuencias (para que la cuasiespecie pueda mantener su identidad), propone el modelo de los *hiperciclos*, un conjunto autocatalítico de replicadores (cada uno de ellos representado por una cuasiespecie), que pudo aportar la masa crítica necesaria de información genética para servir de punto de partida para la vida celular. Eigen sugiere incluso que los primeros tRNA pudieron evolucionar como miembros de un conjunto hipercíclico, que además incluía formas de mRNA; en esto sigue las ideas de [Crick et al., 1976] que postula formas de traducción de mRNA hacia péptidos sin necesidad de actividad enzimática. La aproximación de Eigen ha sido actualizada y desarrollada por E. Szathmáry, J. Maynard Smith y otros [Szathmáry y Demeter, 1987; Maynard Smith y Szathmáry, 1995; Zintzaras et al., 2002].

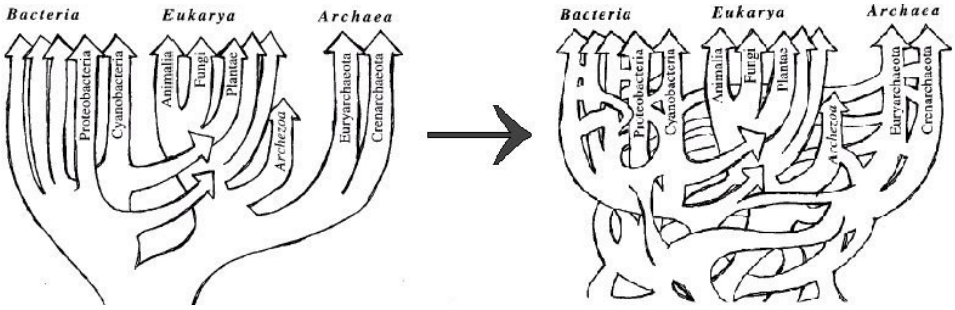


Figura 3.3: La transferencia horizontal dificulta encontrar las raíces del árbol de la vida.

3.2.3. La búsqueda de evidencias

¿En qué basarse para decidir entre las diversas teorías? En primer lugar, en la información que se tiene sobre las características de la Tierra primitiva. Por desgracia, como hemos visto más arriba, existen dudas sobre el tema. En segundo lugar, es importante saber

qué reacciones pudieron ocurrir, dadas esas características. Existe mucha discusión sobre eso también, pero al menos en ese caso el trabajo experimental ha ido aportando algunas respuestas concluyentes sobre algunos puntos, y sin duda aportará más en el futuro, incluso cercano. La otra fuente de información son los organismos actuales: mediante estudios filogenéticos se ha buscado determinar las características de *LUCA* (*Last Universal Common Ancestor*, el Último Ancestro Común Universal). Por desgracia, ya está claro que LUCA, de existir, debió ser bastante complejo (ya contaba con ribosomas, y por lo tanto, con toda la maquinaria de traducción del material genético), por lo que no ayuda a elucidar el origen de la vida; además, ha sido situado como cercano a los procariotas, pero también como un eucariota (a partir del cual los procariotas habrían evolucionado por simplificación), y también se ha discutido acaso es posible extraer alguna información en absoluto a partir de la filogenia tratándose de tiempos tan lejanos [Penny et al., 2003].

Por si esto fuera poco, en los últimos años los proyectos de secuenciamiento han revelado que ha existido una gran *transferencia horizontal* de genes entre especies e incluso reinos distintos, sobre todo en las formas de vida más simples, con lo que el árbol se ha desdibujado hasta hacerse casi indiscernible (la Figura 3.3 fue tomada de [Doolittle, 1999]), y ha debido repensarse todo lo que se sabía sobre la evolución de los procariotas [Gogarten et al., 2002]. Incluso C. Woese [Woese, 2002; Maher, 2002] ha propuesto deshacerse de la idea de una raíz única, sugiriendo que la porosidad de las células primitivas permitió un amplio intercambio de material genético entre los ancestros de lo que luego serían los tres reinos, Eubacteria, Archaea y Eukarya. Textualmente señala: "La transferencia horizontal fue la principal fuerza evolutiva. Las células primitivas no tenían una traza genealógica estable. La evolución celular primitiva es básicamente comunitaria". Recién al alcanzarse un cierto nivel de complejidad (la "barrera Darwiniana"), la transferencia vertical se habría impuesto como el mecanismo principal de herencia.

3.2.4. Fósiles moleculares

Si no hay forma de trazar la genealogía de los organismos, y menos aún de llevarla hasta el origen mismo, lo que sí es posible es trazar la genealogía de genes y moleculas, así como de los mecanismos principales a través de los que estos interactúan: sin duda las secuencias más antiguas fueron aquellas que intervienen directamente en la síntesis de ácidos nucleicos y proteínas, y muchas de ellas son bastante universales en los seres vivos. Se ha acuñado la expresión de "fósil molecular" para referirse a secuencias conservadas en el árbol de la vida y que con alta probabilidad son extremadamente antiguas; el término puede ser ampliado para incluir también mecanismos, y en este sentido, el propio código genético es un fósil molecular.

No son muchos los fósiles moleculares. Un error frecuente ha sido el suponer que las secuencias presentes en las formas de vida más simples *actualmente existentes* están cerca de las formas originales (se esgrimen, por ejemplo, rasgos de los genomas de eubacterias, arqueobacterias e incluso virus). Sin embargo, hay que recordar que las generaciones transcurridas para estos organismos son muchísimas más que para nosotros (son nuestros

sobrinos lejanos, más que antepasados), y sus genomas han estado expuestos a diversas presiones evolutivas; por lo demás, la omnipresencia de la transferencia horizontal nos impide identificar genes presentes en microorganismos diversos como evidencia de un origen arcaico. En pocas palabras: no todo lo simple es antiguo, incluso si se presenta en grupos lejanos. En un sentido sí es viable buscar en bacterias simples o en virus información sobre el origen de la vida: si un mecanismo simple realiza una función, nos demuestra la *factibilidad* de que así se haya hecho en el pasado. Además, en el caso de algunas estructuras específicas, es posible que en los virus la selección haya conducido al desarrollo de versiones prácticamente minimales, lo que es de interés en sí mismo.

Los indicios de la presencia de verdaderos fósiles moleculares son: ubicuidad en el árbol de la vida, función crucial, y naturaleza "no-óptima"; estos elementos sugieren que la molécula se ha conservado por lo crucial de su función -que dificulta la evolución- y no ha sido el resultado de una evolución optimizadora más tardía. Al respecto, véase el apéndice en [Poole et al., 1999], y también [Jeffares et al., 1998].

Todos los fósiles moleculares propuestos según esos criterios son RNA: algunos intrones (en particular los de grupo I y grupo II), algunas ribozimas (en particular las presentes en intrones, así como la RNAsa P), el ribosoma, los snRNAs y snoRNAs (pequeños RNAs, algunos esenciales para la construcción del ribosoma) y sobre todo, el tRNA (o algunas de sus partes). En el caso de las ribozimas, se trata de algunos tipos que aparecen de manera esporádica en el árbol de la vida, aunque también se ha argüido [Salehi-Ashtiani y Szostak, 2001], en el caso de la más conocida y extendida (la ribozima "cabeza de martillo", *hammerhead*) que puede tratarse de evolución convergente (sospecha que siempre hay que tener y chequear cuando se habla de fósiles moleculares). En el caso de los intrones, aceptarlos como fósiles pasa por la aceptación de la teoría sobre su origen antiguo; en los procariotas, presiones evolutivas posteriores los habrían eliminado. Esto no es contradictorio tampoco con la aparente invasión de intrones que se ha producido más recientemente en los eucariotas superiores: muchos intrones pueden ser efectivamente nuevos, y la relajación de algunas restricciones sobre los genomas (que además estaban creciendo) les pueden haber permitido la invasión. Una razón clave para creer en la antigüedad de los intrones es el hecho de que muchos *snoRNAs* son codificados por intrones [Poole et al., 1999; Jackson et al., 2002].

3.2.5. El tRNA como fósil molecular

En el caso del tRNA, se trata de una estructura altamente conservada, esencial, y situada en un punto clave de la máquina celular, pues es el puente entre el RNA y las proteínas; por lo demás, toda teoría del origen de la vida debe pasar por una teoría del origen del código genético. Por estas razones, muchos modelos del origen de la vida lo tienen como protagonista. Como dice el título de [Schimmel, 1996], "el origen del código genético es una aguja en el pajar de las secuencias de tRNA".

La parte más invariante del tRNA (aparte de su forma general) son sus loops, y en ge-

neral, sus *partes libres* (no pareadas). En la Figura 3.4 se detallan las bases conservadas; para establecerlas, reuní información contenida en [Singer y Berg, 1992], [Lewin, 2002] y [Bass, 2001]. La numeración de los nucleótidos corresponde a una convención universalmente establecida. Al igual que en el caso de otros candidatos a fósiles moleculares, cabe preguntarse acaso la repetición de bases en los loops no puede corresponder más bien a evolución convergente, debida a alguna influencia no conocida, en lugar de ser residual. Sin embargo, de ser así, esto debiera notarse en las filogenias que se han construido, y al no encontrar referencia en ese sentido en los artículos que conocemos, damos por hecho que no ha sido el caso, y asumiremos -como todos los autores lo han hecho- que se trata efectivamente de bases preservadas desde ancestros comunes (pero hago la salvedad, pues también es posible que nadie se haya planteado la pregunta al construir las filogenias).

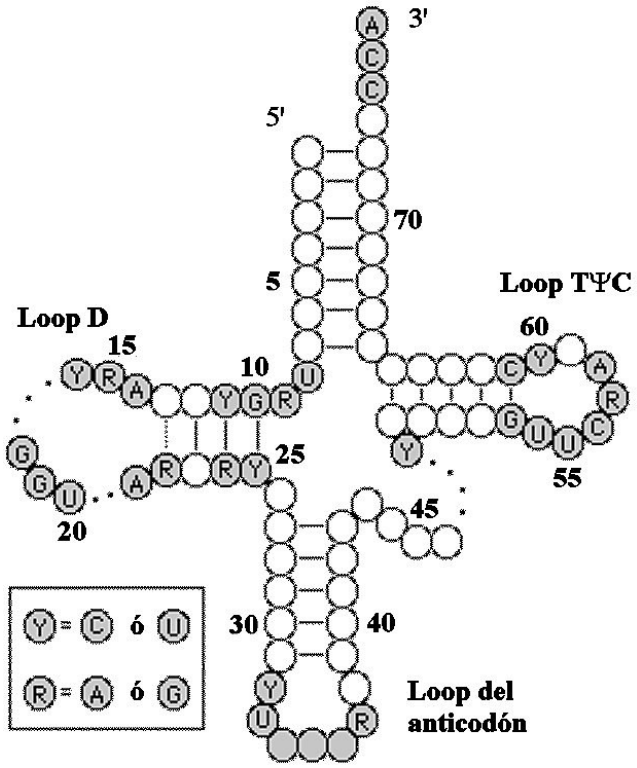


Figura 3.4: Posiciones conservadas en el tRNA.

El origen y evolución posterior del tRNA ha sido tema de muchas propuestas e investigaciones. Respecto al origen mismo, se coincide en suponer que el proto-tRNA consistía en un *hairpin* (un loop con un tallo). Sin embargo, el paso de ahí a la forma actual es controversial: se ha dicho que la parte más antigua son los loops laterales [Maizels et al., 1999], que lo son el loop TΨC y el del anticodon [Maizels y Weiner, 1995], que lo son el tallo y el loop del anticodón [Rodin et al., 1993; Rodin et al., 1996], o incluso que dos hairpins se unieron, formando las mitades derecha e izquierda [Tanaka y Kikuchi, 2001]; naturalmente, cada una de estas ideas mutuamente contradictorias parece tener apoyo en los débiles estudios de comparación de secuencias, y están acompañadas por propuestas de escenarios

que las hacen verosímiles. Ni siquiera hay acuerdo sobre la filogenia de los tRNA, un tema difícil, pues son secuencias relativamente cortas y cuyo árbol familiar debe considerar tanto la gran familia de los seres vivos, como la propia familia de los tRNA (existen docenas, de acuerdo a los distintos anticodones). Se han construido varias filogenias, que al menos demuestran (por ejemplo, [Fitch y Upper, 1988]) que existe necesariamente algún tipo de correlación con el código genético. Para otros el origen no es monofilético [Di Giulio, 1999] (es decir, no hay un ancestro común a todos), cosa que habría que tomar en cuenta para las filogenias. El grupo de investigación de Saks [SaksLab, 2003] desarrolla actualmente amplias investigaciones en torno a la evolución del sistema de traducción, incluyendo el tRNA.

Un punto de especial atención ha sido la aminoacilación, el proceso mediante el cual el tRNA se liga a un aminoácido, y que es catalizado por las aminoacyl-tRNA sintetasas (proteínas). Por un lado, se han estudiado sustratos minimales que responden a esta actividad [Francklyn y Schimmel, 1990; Hipps et al., 1995; Larkin et al., 2002]; por otro, se ha estudiado la forma en que el proceso pudo darse sin necesidad de proteínas, logrando realizar aminoacilación mediante dominios semejantes a tRNA presentes en virus [Felden et al., 1998], mediante una ribozima evolucionada en laboratorio [Lee et al., 2000], e incluso un pre-tRNA evolucionado en laboratorio [Saito et al., 2001], lo que demuestra que una comunidad de tRNAs pudo haber logrado incorporar esta actividad. También se ha estudiado la filogenia de las propias aminoacil-tRNA sintetasas, pues por su función se infiere una gran antigüedad; de particular interés es el hecho de que existen en dos familias, con filogenias independientes, distribuídas ambas a lo largo y ancho tanto del código como del árbol de la vida. El grupo de Schimmel y sus colaboradores parece ser el principal en estas materias, actualmente.

Otro punto por aclarar en cualquier esquema es la función para la cual los tRNA fueron inicialmente seleccionados, y la forma en que de ahí se pasó a la maquinaria celular en la forma que la conocemos. Se ha propuesto, por ejemplo, que el tRNA nació como una "marca" para identificar secuencias e iniciar su replicación (sería un requisito para la evolución de una especie, pues el sistema de replicadores debe ser *cerrado*): es la *genomic tag hypothesis* [Weiner y Maizels, 1987; Maizels y Weiner, 1995; Weiner y Maizels, 1999; Maizels et al., 1999]. Uno de los artículos tiene el sugerente título "Phylogeny from Function: The Origin of tRNA Is in Replication, not Translation". En general, la mirada está puesta en comunidades de estructuras similares a tRNAs, que habrían cumplido simultaneamente funciones enzimáticas y codificadoras, hasta producirse la especialización y la integración de proteínas. También [Ohnishi, 2000] y [Schimmel y Henderson, 1994] tratan el tema general.

3.3. Teorías sobre el código genético

Algunas teorías del origen de la vida tratan de explicar el nacimiento de estructuras autocatalíticas o autopoiéticas a partir de la química existente en los inicios del planeta;

sin embargo, si se quiere explicar el origen de la vida *como la conocemos*, debe explicarse la aparición de la maquinaria celular, y eso requiere una teoría del código genético. El origen del código, por lo tanto, ha sido tema de una infinidad de hipótesis, con distintos grados de base empírica, incluso desde antes de que el código fuese siquiera determinado en los años 60. Incluso más que en el caso de las secciones anteriores, debo advertir que en este campo la literatura es vasta y muy variada (podría escribirse una larga bibliografía usando exclusivamente artículos que incluyan "the origin of the genetic code" en su título), por lo que la exposición es necesariamente simplista y probablemente sesgada.

El código genético es universal, por lo que su origen puede ubicarse en el nacimiento mismo de la vida celular. Las excepciones existentes en algunos organismos han sido satisfactoriamente explicadas como modificaciones posteriores en algunos linajes, por el modelo de "captura de codones" [Osawa et al., 1992; Osawa y Jukes, 1995; Yokobori et al., 2001], a través de una combinación de minimización del genoma y sesgos mutacionales [Andersson y Kurland, 1991], a través de la existencia de mecanismos intermedios con traducción ambigua [Schultz y Yarus, 1994], o mediante una combinación de todos estos [Knight et al., 2001b]. Un review se encuentra en [Knight et al., 2001a]. Una voz disidente es [Syvanen, 2002], que sugiere que el código al principio no fue completamente universal, y fue unificado por una presión evolutiva producto de la transferencia horizontal.

Las teorías se pueden clasificar, a *grosso modo*, en los siguientes tipos:

- **de accidente congelado**
- **estereoquímicas**
- **adaptativas**
- **biosintéticas**

Naturalmente, también se han propuesto casi todas las combinaciones de estas aproximaciones, pues muchos datos pueden ser vistos a la luz de más de una de ellas (por ejemplo: aminoácidos similares pueden ser afines a codones similares; aminoácidos similares pueden estar en la misma cadena biosintética, etc.). Un grupo muy activo en el tema es el de L. Landweber y sus colaboradores, que en [Knight et al., 1999] hacen un review de las principales teorías.

3.3.1. El accidente congelado

La idea del accidente congelado fue propuesta originalmente como una posibilidad a tener en cuenta por el propio F. Crick [Crick, 1968], y según ella el código sería producto de asignaciones arbitrarias, conservadas luego debido a lo crucial de la función que cumple. A estas alturas hay suficientes argumentos para descartarla; si bien algunas asignaciones pudieron ser arbitrarias, la estructura misma del código muestra demasiada correlación con algunas propiedades de los aminoácidos como para ser aleatoria.

3.3.2. Teoría estereoquímica

Se postula que habrían afinidades directas entre los aminoácidos y los codones que los codifican (o los anticodones, o pares codón/anticodón, según la versión de la teoría). Es de especial interés para nosotros, pues la hipótesis de J. Demongeot que se expondrá más adelante parte de una versión fuerte de ella (es decir, atribuyéndole la totalidad del código genético, a diferencia de versiones más débiles que sólo le atribuyen algunas -posiblemente las primeras- asignaciones).

Un temprano proponente de esta idea fue Woese [Woese, 1965; Woese, 1967], y también Crick, quien junto con sugerir la posibilidad del accidente congelado, señaló que era "esencial estudiar la teoría estereoquímica". Ya en 1966 [Pelc y Welton, 1966] propone la existencia de afinidad con los codones, y [Dunnill, 1966] con los anticodones. Versiones posteriores de la teoría son planteadas por [Porschke, 1985] y por [Grafstein, 1983], que propone afinidad con el par codón/anticodón, al igual que lo hacen más tarde [Hendry et al., 1981; Hendry et al., 1995] y [Alberti, 1997]. [Harris et al., 1999] ha sido un partidario reciente de la explicación estereoquímica, que en alguna medida es recogida también en los trabajos de E. Szathmáry [Szathmáry, 1993; Szathmáry, 1999; Zintzaras et al., 2002].

Los artículos de Chipens [Chipens, 1991a; Chipens, 1991b] (por desgracia ambos en ruso) plantean una idea que puede ser relevante para la teoría estereoquímica (ya que sólo conozco los abstracts, ignoro si el autor se pronunció sobre esa relación). Basándose en la idea de Mekler de relación entre aminoácidos y "anti-aminoácidos" (ver Sección 3.2.1), postula la existencia de un algoritmo de codificación que incluye las etapas aminoácido - codón - anticodón - antiaminoácido.

Una ventaja de la teoría estereoquímica es que puede ser evaluada a base de datos experimentales; desafortunadamente, a pesar de la temprana frase de Crick citada arriba, aún no existen resultados concluyentes, y los que hay (así como las proposiciones mencionadas antes) apuntan a afinidades con diversas cosas. El reciente auge de la evolución de RNA *in vitro* (método "SELEX"), si bien ha estado orientado principalmente a la generación de "aptámeros" con fines farmacéuticos, ha generado herramientas que deberían permitir, en pocos años, juzgar definitivamente la viabilidad de esta teoría (con la única salvedad de que tampoco se sabe con certeza cuáles eran las condiciones de la tierra primitiva en que las afinidades debieron darse). [Weber y Lacey, 1978] y [Hobish et al., 1995] encontraron afinidad entre algunos anticodones y sus aminoácidos; más recientemente, se ha demostrado afinidad entre los aminoácidos arginina [Knight y Landweber, 1998; Yarus, 2000] e isoleucina [Majerfeld y Yarus, 1998] con sus respectivos codones. Algunos comentarios en este sentido aparecen en [Landweber, 1999] y en un review [Knight y Landweber, 2000] del estado del tema en el año 2000, y en general sugieren afinidad con los codones (lo que también aparece en el trabajo de Mellersh, citado antes).

También hay quienes han negado por completo que esta teoría tenga base: es el caso de [Strazewski, 1994], que cita un artículo [Mathews, 1988] con el categórico título "No code for recognition".

3.3.3. Teorías adaptativas

Según estas propuestas, habría existido selección entre códigos alternativos, y el que tenemos habría sido el óptimo con respecto a algún criterio. Existen diversas versiones, dependiendo de qué es lo que se supone que se ha optimizado.

[Gilis et al., 2001] usa estadísticas de uso de aminoácidos para mostrar que el código optimiza la estabilidad de las proteínas y se ajusta a la frecuencia de los aminoácidos. [Chechetkin, 2003] señala que el código es óptimo contra mutaciones puntuales y cambios en el marco de lectura.

En [Luo y Li, 2002a; Luo y Li, 2002b] se construye una función global para medir el deterioro por mutaciones, y se concluye también que el código es óptimo para reducir errores; [Haig y Hurst, 1991] y [Labouygues y Figureau, 1984] dan argumentos en el mismo sentido. [Ardell, 1998] argumenta en la misma dirección, de manera bastante sólida, y en [Ardell y Sella, 2001; Sella y Ardell, 2002] combate la principal objeción que enfrenta la teoría adaptativa (que apunta a la dificultad de imaginar un escenario en que códigos alternativos compitan) planteando y simulando un modelo de coevolución entre el código y los genes; el espíritu de [Wills, 1993] es similar.

[Krakauer y Jansen, 2002] proponen un curioso modelo en que el código genético evoluciona como línea de defensa contra parásitos moleculares, para evitar la llegada de elementos genéticos transponibles por la vía de la transferencia horizontal.

Otros han planteado que el código optimiza la "evolvabilidad" de las especies. Éste es el caso de [Kargupta, 2000], y también de [Freeland et al., 2000; Freeland, 2001], el cual en [Freeland, 2001] dice que el código no sólo minimiza errores, sino que además es optimo para adaptarse, al maximizar la conexidad de los paisajes adaptativos (*fitness landscapes*) del espacio de las proteínas; [Aita et al., 2000] y [Dufton, 1997] apuntan en direcciones similares a ésta.

Finalmente, un atípico e interesante trabajo chileno [Soto y Tohá, 1985] postulaba un principio de *hardware minimal* para la codificación de información genética, en que se obtenían óptimos para codones de largo 1, 2, 3 y 4 para 3, 7, 20 y 55 aminoácidos respectivamente; además, deducían que en el caso de 3 y 20, el alfabeto de 4 letras resultaba óptimo. Tohá elaboró posteriormente sobre esto [Tohá et al., 1989], proponiendo un mecanismo para el progresivo crecimiento de los codones.

3.3.4. Teorías biosintéticas

La idea es que al comienzo habría existido un código más degenerado con pocos aminoácidos, y luego los nuevos aminoácidos habrían ido tomando codones de los aminoácidos que los preceden en la cadena biosintética.

El primer y principal exponente de la teoría biosintética es J. Wong [Wong, 1980; Wong, 1988], pero ha sido seguido por muchos otros; una exposición más reciente y contundente

ha sido dada en [Trifonov, 2000]. Una refutación igualmente contundente de los argumentos que soportan la teoría (pero no de la teoría misma, la que no es descartada) se encuentra en [Ronneberg et al., 2000]; también [Aminorvin, 1997] refuta la fuerza de las evidencias estadísticas, y [Di Giulio y Medugno, 2000] le contesta.

Recientemente [Stevenson, 2002] combina la teoría de Wong con la aproximación de Szathmáry, en un modelo de co-evolución del código genético con la replicación de ribozimas.

3.3.5. Código en expansión, código RNY y códigos circulares

Un punto de vista que se combina sin dificultad con varias de las teorías es aquel que plantea un código genético primitivo más degenerado, eventualmente con menos discernimiento de las bases, que habría codificado menos aminoácidos, y que luego habría sido refinado en el tiempo según uno u otro criterio, agregando aminoácidos o precisando su traducción. [Hartman, 1995], por ejemplo, se apoya en la evolución de las aminoacyl-tRNA sintetasas para proponer un código original en que **gg**, **cg**, **gc** y **cc** codifican a los cuatro aminoácidos más simples. [Fitch y Upper, 1988] propone una reducción progresiva de ambigüedad, [Seligmann y Amzallag, 2002] un refinamiento de la codificación de los aminoácidos en tres sucesivos niveles de especificidad química, y [Brooks et al., 2002] una incorporación progresiva de aminoácidos menos abundantes.

Muchos han planteado un código original basado sólo en las letras R (purinas, **g** y **a**) e Y (pirimidinas, **c** y **u**), con **g** y **c** como sus representantes más frecuentes. Ya Eigen [Eigen y Schuster, 1978] proponía los dos codones complementarios **ggc** y **gcc** como los más antiguos, codificando glicina y alanina, respectivamente, los dos aminoácidos más simples. Más en general, Eigen propuso un código original basado en el patrón RNY (donde N es cualquiera de las cuatro bases); otros han propuesto RRY (Crick) y otros patrones. La inspiración de estas ideas está en consideraciones sobre los posibles mecanismos primitivos, en la estructura actual del código, en la sobrerrepresentación del patrón RNY en el loop del anticodón en los tRNA, y en general en su sobrerrepresentación (o la de los otros patrones propuestos) en los genes de los organismos contemporáneos; al menos para esto último, se pueden plantear otras posibles explicaciones, no residuales.

Una ventaja de un código basado en RNY es que se trata de un "código sin coma": no se necesita una marca que indique cómo cortar los codones, pues hay sólo una forma de hacerlo que se ajusta al código. Además, la secuencia complementaria también resulta ser de la forma RNY, permitiendo así codificar información en ambos sentidos [Konecny et al., 1995]. Algunos suponen, por lo tanto, que los tRNA con anticodones de forma RNY serían originarios, y que los aminoácidos así codificados fueron los que primero se utilizaron. Una crítica de esta teoría, o al menos de las supuestas evidencias estadísticas, es [Jukes, 1996].

En una línea parecida, D. Arqués y C. Michel [Arquès y Michel, 1995; Arquès y Michel, 1996] han propuesto la existencia de un *código circular*, también autocomplementario, vestigios de cuya presencia identifican estadísticamente en las secuencias actuales y que

incluso han utilizado con cierta eficacia para ubicar genes en genomas [Arquès et al., 2002]. Dos artículos que básicamente comentan esta teoría son [Pirillo, 2001; Jolivet y Rothen, 2001]; también hay un comentario, con cálculos que repiten los de Arqués pero llegan a resultados bastante menos concluyentes, en [Moreira, 1999].

Capítulo 4

Ideas de J. Demongeot

4.1. Cadena maximal para un código no degenerado

El código genético es degenerado (pues 64 codones se mapean a 20 aminoácidos y una señal de término). Más de alguien se ha preguntado acaso este código puede provenir de un código primitivo no degenerado. En 1983 Jacques Demongeot [Demongeot y Besson, 1983] propuso una cadena maximal para un código semejante, del tipo "cíclico acortado" (*shortened cyclic code*). La misma idea ya la había adelantado en [Demongeot, 1978] e incluso en un apéndice a su Tesis [Demongeot, 1975], aunque el tema general de ésta era otro. Primero son necesarias algunas definiciones.

4.1.1. Definiciones

Sea B un conjunto finito de k elementos (alfabeto); podemos suponerlo de la forma $B = \{0, \dots, k-1\}$. Un código k-ario de orden n construido sobre B es un subconjunto $C \subset B^n$; la función $f : C \to A$ de codificación envía cada elemento de C a un alfabeto A (en el que está escrito el mensaje que se codifica). Si f es biyectiva, decimos que el código es no degenerado; si es sobreyectiva pero no inyectiva, se dice que el código es degenerado o redundante.

El código se dirá *de encadenamiento* (generalizando la definición de *shortened cyclic code*) si existen una matriz S de dimensiones $n \times p$ (con $n \leq p$), una matriz cuadrada T de tamaño p y un elemento $y \in B^p$ tales que

$$\forall x \in C, \qquad \exists\, i \in \mathbb{N} \ \text{ tal que } \ x \,=\, S\,T^i\,y$$

Sea $i_0 = \inf\{i : f(ST^i y) \in A\}$; se llama *ciclo* del código C al entero r definido por

$$r \ = \ \inf\{i > i_0 : T^i y = T^{i_0} y\} - i_0$$

Naturalmente, siempre se tendrá $r \geq \#A$.

Una *cadena maximal* para el código C es un elemento de A^r en la que el i-ésimo signo corresponde al primer signo de $ST^{i_0+i-1}y$.

4.1.2. Cadenas maximales para el código genético

En este caso k será 4 y B corresponderá a las cuatro bases del RNA, representando **u**, **c**, **a** y **g** por 0, 1, 2 y 3, respectivamente. Por otro lado, Demongeot considera que se deben codificar 22 caracteres en A: los 20 aminoácidos, sumados a la señal de inicio y a la señal de término. Por lo tanto, $p = 22$. La matriz T será la matriz de permutación circular

$$T = \begin{bmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ & \vdots & & & & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \\ 1 & 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

mientras que S será una matriz de proyección:

$$S = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \end{bmatrix}$$

y el vector y será una cadena maximal (así que $r = p = 22$). De este modo, los elementos de C se obtendrán leyendo la cadena y en forma circular, como se ve en la Figura 4.1.

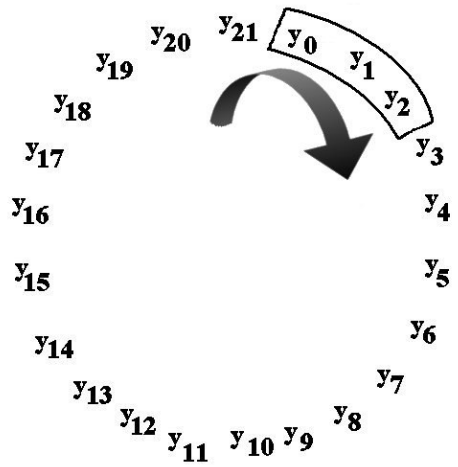


Figura 4.1: Lectura de la cadena maximal.

Demongeot propone las cadenas maximales

$$y_1 = (20330311200123222032 10), \qquad y_2 = (20330311200123210 2032)$$

determinadas por simple inspección. La primera la propone en [Demongeot, 1975] y la recuerda en [Demongeot, 1978]; la segunda, la propone en [Demongeot, 1978] y la retoma en [Demongeot y Besson, 1983]. Ambas satisfacen la misma propiedad: al evaluar la función f, determinada por el código genético, sobre las 22 tripletas, se obtiene una y sólo una vez cada uno de los aminoácidos, además de la señal de inicio y la señal de término.

Aquí hay que hacer una salvedad, que le quita algo de elegancia a esta aproximación: "la señal de inicio" es precisamente uno de los 20 aminoácidos, y más aún, es un aminoácido para el cual existe un sólo codón codificante (en el código genético completo), la metionina. Por lo tanto, el codón **aug** (203) aparece dos veces en cada una de estas cadenas, y la decisión de traducirlo una vez como metionina y la otra vez como "inicio" resulta arbitraria.

Cabe hacer notar que, con $p = 64$ y matrices S y T apropiadas, es posible escribir una cadena maximal para los 64 codones del código genético; la existencia de dicha cadena deriva de la existencia de ciclos Eulerianos en el grafo de De Bruijn. Una observación interesante es que, como se hace notar en [Demongeot y Besson, 1996], la cadena y_2 puede ser extendida a uno de esos ciclos Eulerianos:

213332331212301002111**0322033031120012232102**011300030231013132022

Otra observación que se hace en [Demongeot y Besson, 1983] es un comentario al artículo [Cullman y Labouygues, 1982], en el que se daba un código de Baudot, binario, de orden 5, con cadena de largo 32, para cubrir (de manera completa, aunque con repeticiones) todos los aminoácidos del código genético. Al respecto, Demongeot muestra una cadena de largo 29 que lo hace también, con código de orden 6 (las sextetas binarias se pasan al alfabeto de 4 letras reemplazando 00, 01, 10, y 11 por sus valores decimales). También es degenerado, pero al menos tiene menos bits que el de Cullman, y al igual que en aquel, la degeneración se corresponde (en cuanto a número de aminoácidos repetidos) con la del código genético.

Lo expuesto hasta aquí es el contenido de los artículos antiguos de J. Demongeot; no hay referencia a la forma de generación del código, al contexto biológico en que debio evolucionar hacia formas degeneradas, ni tampoco se propone que esta cadena maximal haya correspondido a estructura biológica alguna. El propio autor señala que "es muy difícil dar un significado biológico a nuestras observaciones", y solamente concluye de ellas que es posible reducir considerablemente la redundancia del código genético, al tomar sólo una subcadena relativamente corta (22 de 64) de una cadena que genera el código completo.

4.2. Un RNA minimal en el origen de la vida

En [Demongeot y Besson, 1996] Jacques Demongeot parte desde una perspectiva completamente distinta, para llegar pronto al problema de un anillo que codifique todos los

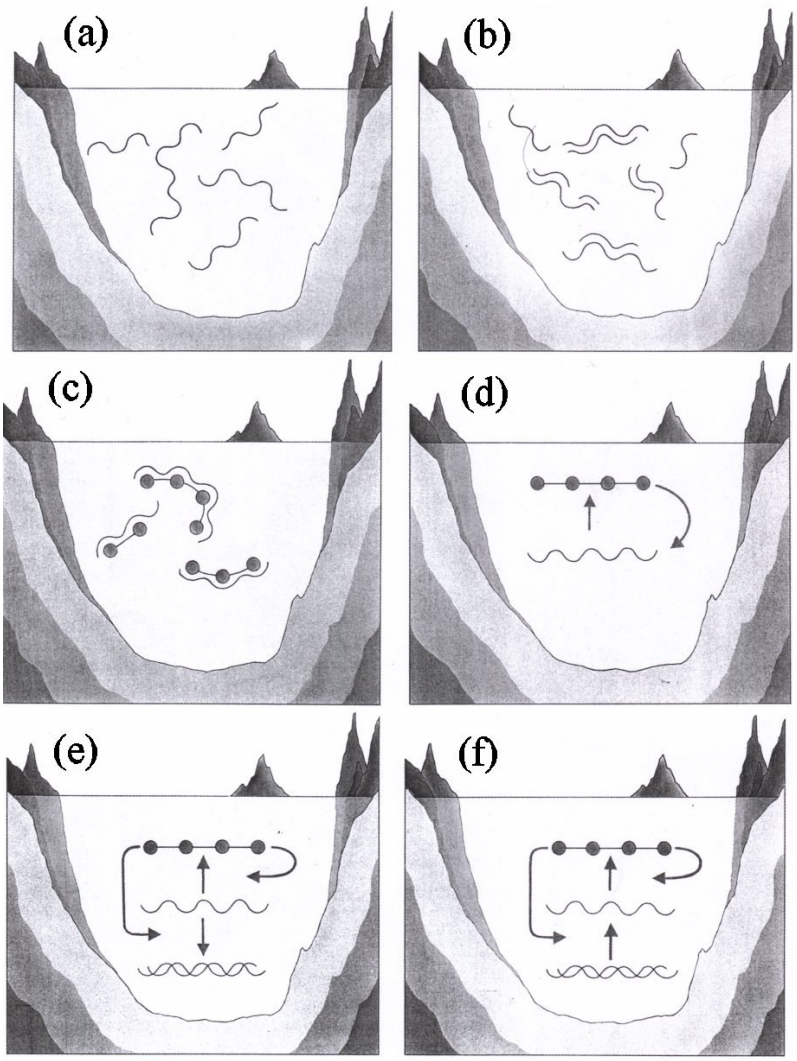


Figura 4.2: Un escenario para el origen de la vida, bastante acorde con el propuesto por Demongeot: (a) las moléculas de RNA aparecen espontáneamente; (b) algunas moléculas de RNA se replican; (c) otras se adhieren a cadenas de aminoácidos; (d) algunas proteínas "mejoran" las cadenas de RNA; (e) las proteínas transforman el RNA en DNA y ayudan al RNA a hacer proteínas; (f) las proteínas controlan la expresión del DNA. Fuente: [Leroy, 1993].

codones.

En primer lugar, recoge la discusión existente sobre los orígenes del código genético, y de los tRNA en particular. Por lo general las propuestas sobre ese tema suponen la existencia de RNAs, en forma de hairpin, o eventualmente en forma de anillo, relacionados con aminoácidos, que luego al fusionarse habrían dado origen a la forma de trébol del tRNA que conocemos. Recoge además la teoría estereoquímica del código genético (puntualmente [Hobish et al., 1995]), y considera que las secuencias de RNA pudieron haberse estabilizado al asociarse con los aminoácidos afines a sus tripletas.

Se sitúa en un escenario bidimensional (como por ejemplo, las capas de arcilla de Cairns-Smith), en las que habrían aparecido las primeras estructuras auto-replicadoras, y plantea un modelo en que un RNA (denotado A en la Figura 4.3), exitoso por algún motivo, junto con su RNA complementario (A^*), habrían interactuado con aminoácidos y contribuido a la formación de las primeras proteínas, mediante -entre otras cosas- la formación de hairpins (tA^*). La Figura 4.3 lista las reacciones químicas que debieron ocurrir entre estas secuencias, hasta la formación de un *pool* de proteínas y RNAs capaces de dar inicio a los mecanismos reproductivos. Las propias secuencias A, A^* y su doble hélice, AA^*, habrían sido a la vez los primeros genes, y los primeros tRNAs (en esto coincide con varias de las otras propuestas existentes).

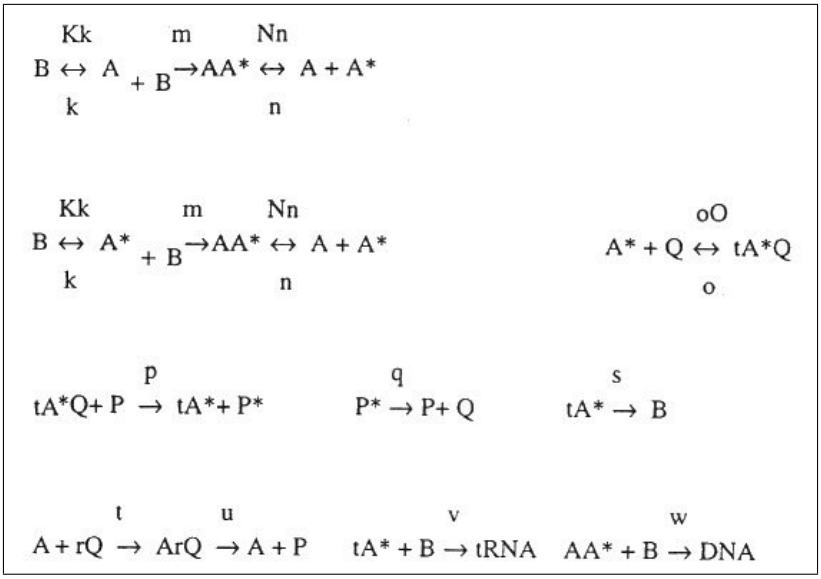


Figura 4.3: Reacciones químicas hipotéticas. B: pool de bases; A: un RNA; A^*: su secuencia complementaria; AA^*: doble hélice de A y A^*; Q: pool de aminoácidos; tA^*: tRNA primitivo; tA^*Q: complejo aminoácido-tRNA primitivo; P: pool de péptidos; P^*: pool de proteínas; ArQ: pool de 'protocélulas'; tRNA: pool de tRNA; DNA: pool de DNA.

Bajo el supuesto de que la afinidad con aminoácidos pudo estabilizar los RNA, y bajo el supuesto simplificador de que los 20 aminoácidos usados por los seres vivos se encontraban presentes en proporciones similares, deduce que habría existido selección a favor de los RNA cuya secuencia fuese afín de manera similar a todos los aminoácidos (ver Figura 4.4). Al pedir además que la secuencia sea lo menor posible, el resultado es pedir que la secuencia A (o equivalentemente, A^*) sea un anillo que codifique una y sólo una vez cada uno de los aminoácidos. Es decir, una cadena maximal al estilo de las presentadas en la sección anterior.

Para satisfacer la codificación de todos los aminoácidos son necesarias al menos 20 letras. Sin embargo, no existen anillos que cumplan esto. Entonces, Demongeot agrega otra condición: que la secuencia se pueda escribir comenzando con una señal de inicio (**aug**) y términando con una señal de término. Como muestra y_2 en la sección anterior, tales

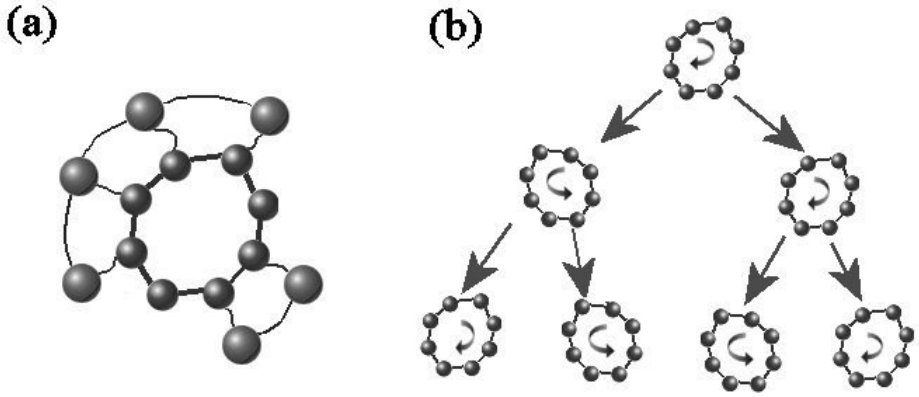


Figura 4.4: El juego de la vida: (a) un RNA circular se estabiliza mediante la afinidad con aminoácidos, con lo que RNA que "aprovechan" mejor el pool de aminoácidos tienden a durar más, y aumentar en población respecto a otros; (b) durando lo suficiente, puede darse replicación mediante pareos Watson-Crick; se alternan el RNA y su forma complementaria.

secuencias existen.

Para encontrar *todas* las secuencias de largo 22 que verifican esta condición (dentro de las $4^{22} \approx 1{,}7 \times 10^{13}$ secuencias posibles), Demongeot y sus colaboradores, como se informa en [Weil et al., 1995], plantearon el problema como un problema de satisfacción de restricciones (*constraint satisfaction problem*), y lo resolvieron usando software ad-hoc disponible comercialmente. El resultado fueron 1280 soluciones, de las cuales 672 finalizan con el codón **uga**, 256 lo hacen con **uaa**, y 352 lo hacen con **uag**. Confirman además que no existen anillos de 20 bases que codifiquen los 20 aminoácidos, y verifican que no existen anillos de largo 21 que codifiquen todos los aminoácidos más *stop* (es decir, que tengan un codón en cada clase sinónima del código genético).

Ya hemos visto en el capítulo anterior que el tRNA es considerado un fósil molecular; en particular, en lo que respecta a sus loops y partes libres. Ya en [Hartman, 1984] se proponía el estudio de la estructura de loops del tRNA, y se sugería que el propio loop del anticodón podía derivar de un tRNA primitivo con forma de hairpin. Por otro lado, se ha argumentado más de una vez, partiendo por el propio Eigen, que los tRNAs más antiguos debieran ser los de la glicina, el aminoácido más simple y el que aparece en mayor cantidad en los experimentos de síntesis bajo condiciones prebióticas. Recogiendo todo esto, Demongeot procedió a comparar las partes libres de distintos Gly-tRNA disponibles en la literatura del momento con los anillos terminados en **uga** (ignoro por qué sólo esos; tal vez por ser los más numerosos).

Tras hacer esta comparación, escoge un anillo en particular, al que bautiza como "código AB", que tiene la virtud de tener una buena estructura de hairpin (ver Figura 4.5), y además resulta extremadamente similar a las partes libres del Gly-tRNA mitocondrial de

la planta *Œnothera sp.* (ver Figura 4.6). Curiosamente, la secuencia de este anillo,

$$\text{AB} \;=\; (\mathbf{auggugccauucaagacuauga})$$

resulta ser la misma de y_2, la cadena maximal propuesta ya en [Demongeot, 1978], mucho antes de que estuviese disponible la secuencia de los tRNA mitocondriales de *Œnothera*, publicada recién en [Binder et al., 1990].

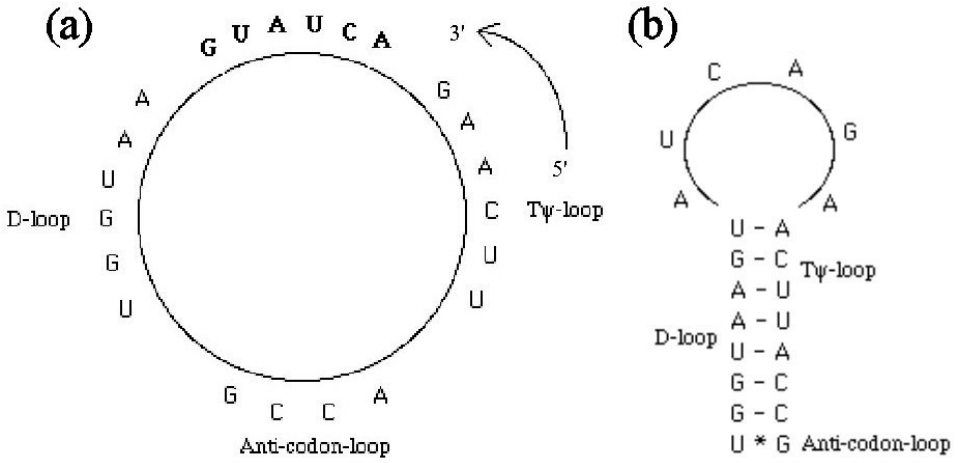


Figura 4.5: El anillo AB: (a) forma circular; (b) hairpin.

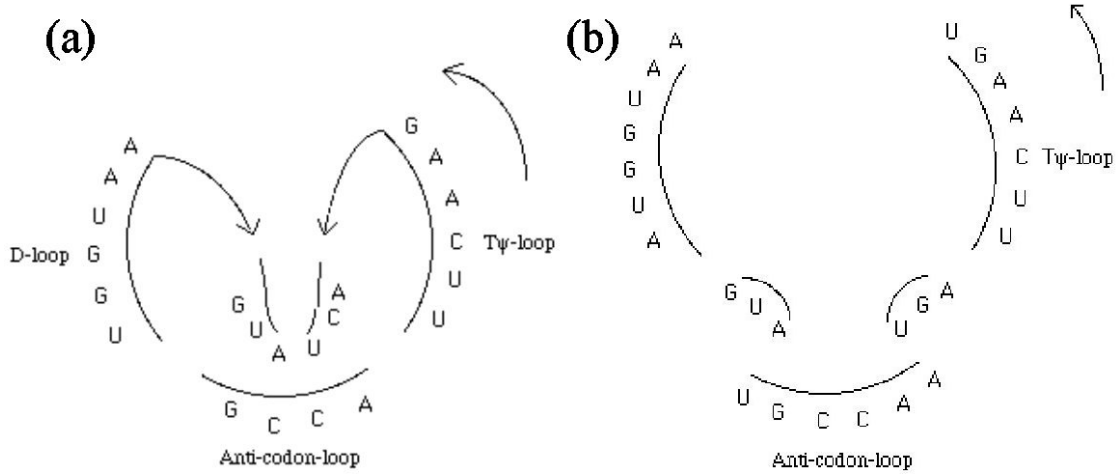


Figura 4.6: Comparación de AB con un fósil molecular: (a) AB denaturado; (b) partes libres del Gly-tRNA mitocondrial de *Œnothera*.

Algunas propiedades adicionales que Demongeot observa en el anillo AB (al que se refiere como "AB code", en consideración a la idea de cadena maximal para un código generado, según lo expuesto en la sección anterior), son las siguientes:

- Excepto por el dinucleótido **cg**, la cadena contiene los otros 15 posibles comienzos de codones, cosa que Demongeot relaciona con la "hipótesis del comodín" (*wobble hypothesis*) de F. Crick.
- El hairpin es similar a las formas de hairpin ancestral propuestas por Di Giulio [Di Giulio, 1995].
- Contiene 6 codones de la forma RNY (de los 8 que podría contener), cosa que Demogeot relaciona con la hipótesis RNY en la forma presentada en [Konecny et al., 1995].
- El anticodón de la glicina queda precedido, en el loop de anticodón de AB, por **u**, al igual que en todos los tRNA contemporáneos.
- En trabajo con un estudiante [Le Touze, 1995] encuentra amplia similitud entre AB y las partes libres de una fracción importante de otros 48 tRNAs encontrados en la literatura.

Además, encuentra similitudes entre otros anillos (de los 672 comparados) con algunos otros tRNA, de otras especies y aminoácidos. Uno de estos, similar a un Ala-tRNA de *Neurospora crassa*, resulta ser además muy similar a la secuencia complementaria de AB, lo que coincide con la idea de [Rodin et al., 1993], respecto a la complementariedad primitiva entre los tRNA asociados a anticodones complementarios.

También compara la lista de codones usados en AB con las frecuencias de uso de codones tabuladas a la fecha, encontrando alguna relación con los codones más usados en algunas especies. Sin embargo, los resultados ahí son poco concluyentes, y en el trabajo de su estudiante G. Le Touze [Le Touze, 1995], con el genoma de *E. coli*, la hipótesis de una relación positiva debió ser rechazada.

Otro análisis es la comparación del contenido de *gc* (cantidad de bases **g** ó **c**) y la redundancia I (una medida de distancia de la secuencia respecto de una secuencia aleatoria), entre AB y los genómas contemporáneos (según datos un poco antiguos). El resultado es el gráfico de la Figura 4.7, donde la curva gruesa representa el límite permitido por la relación entre *gc* e I, que se alcanza para la máxima independencia entre las bases. La observación que hace Demongeot es que a medida que se avanza en la evolución, las especies se van alejando de esta curva, al introducirse secuencias organizadas y repetitivas en el genoma; en tal caso, el código AB queda situado especialmente cerca de la curva, lugar apropiado para un candidato a genoma ancestral.

Finalmente muestra (como ya reprodujimos en la sección anterior) que AB puede incluirse en un anillo de 64 bases que contiene todos los codones (cosa imposible, señala, con los 608 anillos que no terminan en **uga**).

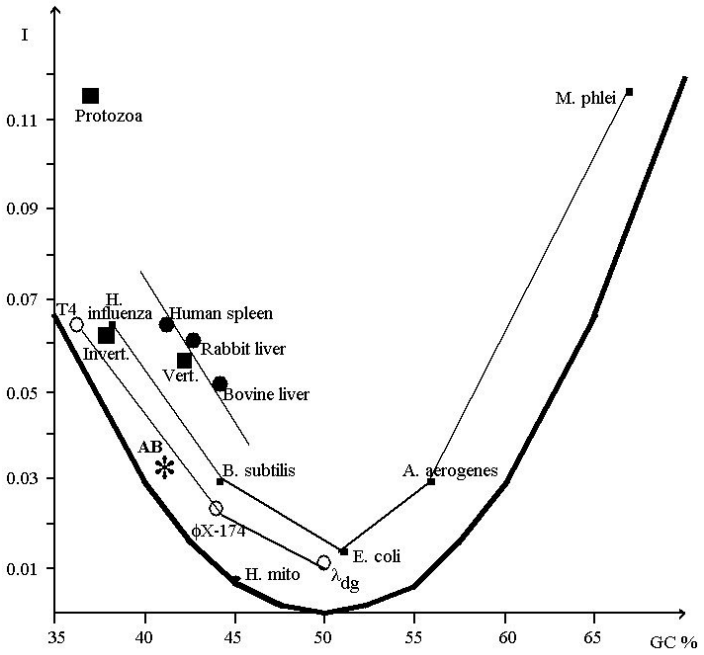


Figura 4.7: AB y algunos genomas: gc vs. I.

4.3. Resumen: supuestos, tesis y posibles debilidades

Resumiendo, lo que Demongeot propone es que existió un juego, en la "pizza" primigenia, en que cadenas cortas de RNA flotaban, se rompían, se volvían a formar, tomaban forma de hairpin o de anillo, se pareaban en pequeñas doble hélices, etc. Al existir la posibilidad de estabilizarse mediante afinidades directas con los aminoácidos presentes en el medio, los anillos ganadores de este juego habrían sido los que correspondían, en los términos de 4.1.2, a cadenas generadoras maximales del código genético. Al listar estos anillos, Demongeot encuentra uno (AB) con una forma de hairpin relativamente buena y que coincide casi por completo con las partes libres de un tRNA contemporáneo (y en forma importante con algunos otros tRNAs), lo que concuerda con la idea de que estos anillos dieron origen a los genes y especialmente al tRNA, mediante la inserción estabilizadora (o diferenciadora de funciones) de las partes lineales; a esto agrega una serie de otras propiedades interesantes.

Es bueno tener presentes los supuestos que avalan todo el razonamiento:

- Se da por sentada la presencia de nucleótidos, los que en los experimentos con química prebiótica, si bien aparecen, son más difíciles de obtener que las bases sueltas, o los aminoácidos.
- Se prescinde de catalizadores de la polimerización de RNA ó péptidos, atribuyéndola, por ejemplo, a las actividades del propio medio, tema aún controversial.

- Se asume una versión especialmente fuerte de la teoría estereoquímica del código genético, considerándolo como una tabla de afinidades de codones con aminoácidos (al menos para los codones utilizados por AB).
- Se asume una distribución pareja de los aminoácidos en el medio. Esta simplificación es bastante discutible, como también lo es la ausencia de otros aminoácidos, no codificados por el ácido genético, pero que aparecen en los experimentos de biosíntesis espontánea. Levantar este supuesto ciertamente cambia "las reglas del juego".

A las posibles debilidades de los supuestos, se agregar algunos otros problemas. Uno especialmente grave es que en la Figura 4.6, al comparar con las partes libres del tRNA, las tripletas **gua** y **uca**, alineadas con **gua** y **uga** del tRNA, quedan *al revés*, es decir, con el sentido $5' \rightarrow 3'$ invertido. Eso es malo, pues la única forma de invertir el sentido de la molécula es separando por completo todas las bases y pegándolas de nuevo; esto invalida esa parte del alineamiento.

A esto se agregan otras preguntas. ¿Por qué pedir anillos que empiecen con **aug** y terminen con un codón de término? Esa es la forma de los genes actuales, pero no hay razón para imponerla en los competidores del juego. Además, ¿qué tiene de especial AB? ¿Algo lo distingue, antes de comparar con los tRNA actuales? ¿Y qué tiene de especial *Œnothera*? ¿Se mantienen las propiedades interesantes, a la luz de los conocimientos actuales de genomas (mucho más masivos y mejor organizados que en 1996?) Estos y otros problemas motivaron una nueva visita a este tema, trabajo emprendido por el autor de la presente Tesis bajo la orientación del profesor Demongeot.

Capítulo 5

Simulaciones

Dentro de las ideas de J. Demongeot expuestas en el capítulo anterior aparece un "juego de la vida", en que nucleótidos y aminoácidos se mueven e interactúan según ciertas reglas en un medio que se asume bidimensional. Se afirma dentro de su propuesta, que el ganador de ese juego será un RNA con forma de anillo, capaz de codificar una y sólo una vez cada uno de los aminoácidos.

Para verificar esto, intentamos simular un modelo simplificado de estas interacciones. El primer intento fue una simulación espacial, esto es, con partículas ubicadas en un plano, moviéndose al azar e interactuando de acuerdo a una versión simple de las reglas. Esta simulación, que revisaremos a continuación, debió ser abandonada por ser computacionalmente demasiado costosa; las instancias necesarias para poder ver alguna evolución tendrían que ser muy grandes y correr durante mucho tiempo, por lo que son inviables. El programa resultante puede ser, en el mejor de los casos, usado para fines demostrativos (para explicar el "juego" que se plantea en la teoría).

Como alternativa, se planteó luego una simulación *no espacial*, con listas de partículas, carentes de coordenadas, a las que se hace interactuar mediante encuentros aleatorios entre distintos índices de la lista. Este modelo, menos realista, tiene la ventaja de ser viable para cantidades mayores de elementos, y permitió chequear algunas de las ideas sobre el resultado del juego, con resultados débiles, pero que apuntan en la dirección esperada. Finalmente, se presentarán algunas ideas con las cuales se espera que la simulación mejore su verosimilitud (y con ello, esperamos, el apoyo a la hipótesis).

5.1. Consideraciones para un código genético simplificado

El principal problema de las simulaciones fue lograr echar a correr un mundo suficientemente grande, y durante suficientes iteraciones, para que fuera posible observar alguno de

los fenómenos que nos interesan. Una forma de hacer esto más viable fue la reducción del número de tipos de bases y aminoácidos, a 2 y 4, respectivamente. Mantuve el tamaño de los codones (3 bases) pues reducirlos no simplificaba mayormente la simulación, y en cambio, hacía más difícil interpretarla.

Se necesitaba, por lo tanto, un código genético con dos tipos de base (0 y 1), 4 aminoácidos (A,B,C,D) y codón de largo 3, por lo que $2^3 = 8$ combinaciones de bases debían mapearse a 4 letras. Existen $4^8 = 65536$ códigos posibles (funciones $\phi : \{0,1\}^3 \to \{A,B,C,D\}$). Pero además, queríamos capturar la situación de la manera más realista posible, lo que incluía contar con un nivel de degeneración similar al del código genético real. En nuestro caso, existen 5 esquemas de degeneración posibles:

1 : 1 : 1 : 5
1 : 1 : 2 : 4
1 : 1 : 3 : 3
1 : 2 : 2 : 3
2 : 2 : 2 : 2

Se escogió $1 : 2 : 2 : 3$, por ser el más cercano al código real (*i.e.*, 1 codón para A, 2 para B, 2 para C y 3 para D). Además, una idea que nos interesaba verificar en las simulaciones era acaso los codones no usados en el anillo "óptimo", formarían anillos suboptimales; por esto, este esquema de degeneración tiene además la ventaja de que una vez eliminados un codón por cada aminoácido, sigue siendo posible codificar tres de ellos.

Podemos determinar el número de posibles ϕ para este esquema: escogemos primero la clase con 1 elemento, luego la de 3, y luego quedan 3 formas de separar los restantes 4 en 2 y 2 (ya que B y C son etiquetas intercambiables). Tenemos entonces un conjunto, que llamaremos $\mathcal{C}$, con cardinal

$$8 \times \binom{7}{3} \times 3 = 840$$

A fin de tener algo semejante a los anillos de Demongeot, necesitamos que existan anillos de largo 4 capaces de codificar los cuatro aminoácidos (y ojalá, por lo dicho arriba, anillos capaces de codificar 3). Descartando rotaciones, los posibles anillos son

$$0000, \quad 0001, \quad 0011, \quad 0101, \quad 0111, \quad 1111$$

Los anillos 0000, 0101, y 1111 deben ser descartados, pues no generan 4 (ni siquiera 3) codones diferentes. 0111 es equivalente a 0001, y por lo tanto no necesita ser considerado. En consecuencia, sólo sirven 0011 y 0001.

0011 tiene la ventaja de tener el mismo número de 0 y 1, y de ser cerrado bajo complementación. El número de posibles ϕ (cuyo conjunto denotaremos $\mathcal{C}_0$) que permitirían a 0011 codificar 4 diferentes aminoácidos (nuevamente, descartando opciones equivalentes bajo $B \leftrightarrow C$) es

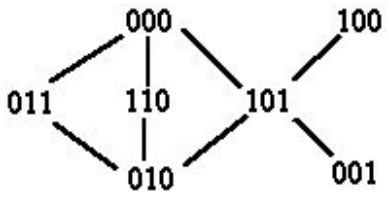
$$4 \times 3 \times \binom{4}{2} \times 2 = 144$$

Una desventaja es que los codones que quedan tras remover los usados en 0011, es decir, {000, 010, 101, 111}, no incluyen 3 capaces de formar un anillo de largo tres (es decir, no podrían existir el "segundo mejor" con las características mencionadas arriba).

0001 no es cerrado bajo complementación (produce el anillo equivalente, 0111); los códigos que le permiten codificar 4 aminoácidos distintos (a los que llamaremos $\mathcal{C}_1$), son al igual que en $\mathcal{C}_0$, 144. Por otro lado, sí es posible en este caso producir un anillo de largo 3 (011) que contiene tres codones no producidos por 0001. El número de posibles ϕ que permiten a 0001 y 011 codificar respectivamente 4 y 3 aminoácidos distintos (a los que llamaremos $\mathcal{C}_2$) es

$$4 \times 3 \times 3 \times 2 = 72$$

Sólo existen 6 códigos ϕ en $\mathcal{C}_0 \cap \mathcal{C}_2$. Para verlo, notemos que los codones dentro de cada uno de los conjuntos siguientes deben codificar cosas distintas: $\{001, 011, 100, 110\}$, $\{000, 001, 010, 100\}$, $\{011, 101, 110\}$. Forzosamente, $\phi(111) = D$. El grafo que relaciona los codones que pueden ser sinónimos está en la figura a continuación.



Para 100, 101 y 001 no pueden haber tres valores distintos, ni tampoco tres iguales; además, $\phi(100) = A \iff \phi(101) = \phi(001)$, e igualmente, $\phi(001) = A \iff \phi(101) = \phi(100)$. Los dos casos son equivalentes, si se revierte el orden de las bases en los codones ($abc \leftrightarrow cba$), por lo que nos quedamos con el primero. Resta solamente definir de quién es sinónimo 000 (110 ó 011), y ver cuál de los pares de sinónimos se asocia a D. El resultado son sólo 6 códigos realmente distintos:

$\phi(000)$	$\phi(001)$	$\phi(010)$	$\phi(011)$	$\phi(100)$	$\phi(101)$	$\phi(110)$	$\phi(111)$
D	A	B	D	C	C	B	D
B	A	D	B	C	C	D	D
B	A	C	B	D	D	C	D
D	A	B	B	C	C	D	D
C	A	B	B	D	D	C	D
B	A	D	D	C	C	B	D

En tablas semejantes a la que se usa para representar el código real, se ven como sigue:

Primera base	Segunda base			
	0		1	
0	000	D	010	B
	001	A	011	D
1	100	C	110	B
	101	C	111	D

Primera base	Segunda base			
	0		1	
0	000	B	010	D
	001	A	011	B
1	100	C	110	D
	101	C	111	D

Primera base	Segunda base			
	0		1	
0	000	B	010	C
	001	A	011	B
1	100	D	110	C
	101	D	111	D

Primera base	Segunda base			
	0		1	
0	000	D	010	B
	001	A	011	B
1	100	C	110	D
	101	C	111	D

Primera base	Segunda base			
	0		1	
0	000	C	010	B
	001	A	011	B
1	100	D	110	C
	101	D	111	D

Primera base	Segunda base			
	0		1	
0	000	B	010	D
	001	A	011	D
1	100	C	110	B
	101	C	111	D

Otro criterio de interés a la hora de escoger el código puede ser el efecto de las mutaciones puntuales; dentro de las teorías adaptativas del código genético, algunas proponen que se escogió para este efecto. En cada uno de los seis códigos anteriores, por ejemplo, existen 6 mutaciones puntuales (cambios de un bit en el codón), de las 24 posibles, que mantienen el aminoácido codificado; están distribuidas en 6 codones, para los dos códigos de la derecha, o concentradas en 4 de ellos, en los demás. El óptimo para el esquema de degeneración que estamos ocupando es 8 (2 en ambas clases formadas por dos codones, y 4 en la formada por tres), y se alcanza, por ejemplo, en los siguientes códigos, pertenecientes a $\mathcal{C}_0$, el de la izquierda, y $\mathcal{C}_1$, el de la derecha:

Primera base	Segunda base			
	0		1	
0	000	C	010	B
	001	A	011	B
1	100	C	110	D
	101	D	111	D

Primera base	Segunda base			
	0		1	
0	000	A	010	C
	001	B	011	B
1	100	D	110	C
	101	D	111	D

Aparte de servirnos para escoger los códigos a usar en las simulaciones, estas consideraciones no están exentas de posible interpretación biológica, considerando las especulaciones existentes sobre un alfabeto inicialmente binario (R/Y, purinas y pirimidinas), y algunas sugerencias (como en [Hartman, 1995]) de que originalmente sólo se habrían codificado los 4 aminoácidos más simples, glicina, prolina, alanina y arginina. Los posibles esquemas de degeneración y el nivel de tolerancia a las mutaciones son relevantes bajo esas hipótesis. Sin embargo, preferimos quedarnos sólo con la idea de un modelo simplificado, sin interpretación, pues bajo esas hipótesis que le darían una interpretación, el modelo de anillos no es viable (pues el RNA tendría que doblarse demasiado). Nuestros anillos de largo 4, por lo tanto, sólo intentan mostrar lo que pudo ocurrir con anillos más largos.

5.2. Simulación espacial

La simulación consistió en un espacio bidimensional, en el que se movían partículas de dos tipos (aminoácidos y bases), con coordenadas continuas. Los aminoácidos estaban caracterizados por una posición, un tipo (0, 1, 2 ó 3), y también podían tener un enlace a una base. Las bases estaban caracterizadas por su posición, un ángulo de giro, su tipo (0 ó 1), y podían estar enlazados a un aminoácido y a una o dos otras bases, por la izquierda y derecha (no alcanzamos a incorporar hibridación antes de desechar el modelo, por lo

que no se usó un cuarto enlace, a base hibridada).

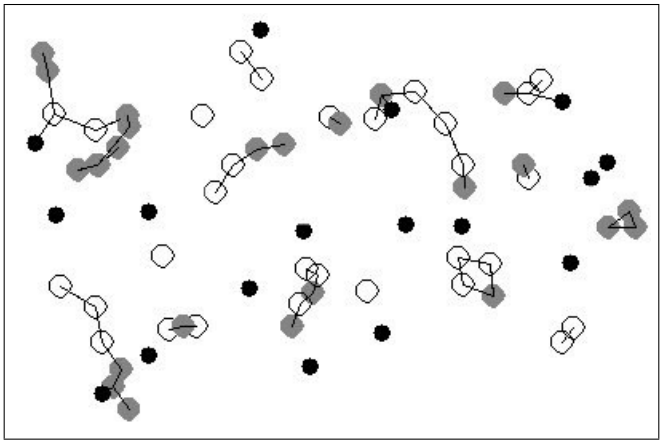


Figura 5.1: Un momento en la simulación espacial. Se pueden ver bases de dos tipos, formando cadenas y anillos. Algunos de los aminoácidos (puntos más pequeños) se han asociado a cadenas.

Tanto para las bases como para los aminoácidos existía un radio representando su volumen (*minBaseDistance* y *minAaDistance*); ningún otro elemento se podía acercar a menos de esa distancia.

El movimiento de cada partícula, en cada iteración, se realizaba sumando a cada coordenada una variable aleatoria uniformemente distribuida en [-*brownianStep*/2,*brownianStep*/2]; en el caso de las bases, también se sumaba una variable aleatoria uniforme en [-0.01, 0.01] a su ángulo de giro (medido en radianes). Si el resultado dejaba a la partícula fuera del área de simulación, o en un espacio ya reservado por otra partícula, entonces el movimiento tampoco se realizaba. En el caso de partículas enlazadas de alguna forma, existía un rango, [*minLinkDistance*, *maxLinkDistance*], dentro del que debía mantenerse la distancia entre los puntos de enlace (también era este el rango en que tenían que encontrarse puntos no enlazados para que el enlace pudiera crearse). El punto de enlace de los aminoácidos eran sus coordenadas, mientras que para las bases existían tres: L (izquierdo), R (derecho) y A (para el aminoácido), a distancia *baseRad* de las coordenadas de la base, y determinados por su ángulo de giro, como indica el esquema de la Figura 5.2.

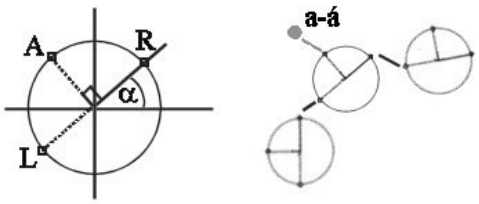


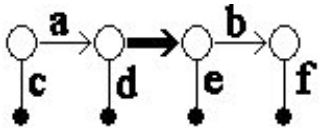
Figura 5.2: Esquema de los enlaces de las bases en la simulación espacial.

En el caso de agregados (componentes conexas de elementos enlazados) se sumaba, a los

intentos de movimiento de cada partícula, un paso aleatorio del agregado completo (con pasos del mismo tamaño). En general, el propósito de introducir el ángulo de giro de las bases, y de permitir un cierto rango de elasticidad en los enlaces, era dar cabida a variaciones en la curvatura de las cadenas, para que de manera natural pudiesen formarse los anillos u otras estructuras (como hairpins, aunque para eso habría sido necesaria la hibridación).

La formación de enlace entre bases estaba regida por un parámetro de probabilidad *probLink*, y requería además que los puntos de enlace (L y R) de las bases estuviesen en el rango de distancias apropiado. La formación de enlace de un aminoácido con una base estaba regido por *probLinkAa*, y aparte de exigir la distancia apropiada, requería que la base formara con sus bases vecinas un codón codificando el aminoácido (de acuerdo a un código genético binario parametrizado en el código); el enlace sólo sobrevivía mientras el codón lo hiciera.

El rompimiento de enlaces entre bases estaba regido por el parámetro *probUnlink*; una vez tomada la decisión de romper un enlace, éste tenía una probabilidad de sobrevivir, dada por una puntuación que modelaba el efecto estabilizador de las estructuras vecinas. Se ensayaron varios esquemas, asignando distintos valores a la existencia de una cadena más larga, y a la presencia de aminoácidos enlazados a las bases, como se indica en el esquema. De este modo, la probabilidad de que el enlace sobreviviera era proporcional a $(a + b + c + d + e + f)$ (normalizando por el valor máximo posible para la suma).



El loop de iteración consistía en mover las partículas individuales, mover los agregados, enlazar bases, enlazar aminoácidos a las bases, y luego tratar de romper enlaces.

Resultados

El código fue escrito en Java, y luego portado a C++ (sin representación gráfica en ese caso), para tratar de hacerlo más veloz. También para mejorar la velocidad se particionó el plano, guardando las partículas de cada sector en una lista enlazada, de modo de chequear distancias sólo con las partículas de los sectores vecinos (y evitar así el crecimiento cuadrático de las comparaciones necesarias).

Aún así, la gran cantidad de trabajo requerido por los cálculos de distancias, ángulos, etc, hizo inviable la simulación a una escala suficientemente grande como para observar los fenómenos que nos interesaban; además, la tasa de formación de anillos y cadenas era bastante baja. Una población de 3.000 partículas, por ejemplo, corrió durante varias horas, por 12.000 iteraciones, sin llegar a producir más de 200 anillos. Y con esa cantidad, repartida entre una quincena de tipos distintos (por forma y tamaño), las poblaciones a través del tiempo oscilaban de manera demasiado fuerte como para notar tendencias claras a favor de un tipo u otro. Ni siquiera fue posible (como sí lo fue luego en la simulación no

espacial) notar el efecto de presiones selectivas tan fuertes como poner sólo aminoácidos de un tipo, codificado por un único codón. Lo único que pudo observarse fue el paulatino aumento de los anillos en desmedro de las cadenas, como muestra el gráfico en la Figura 5.3.

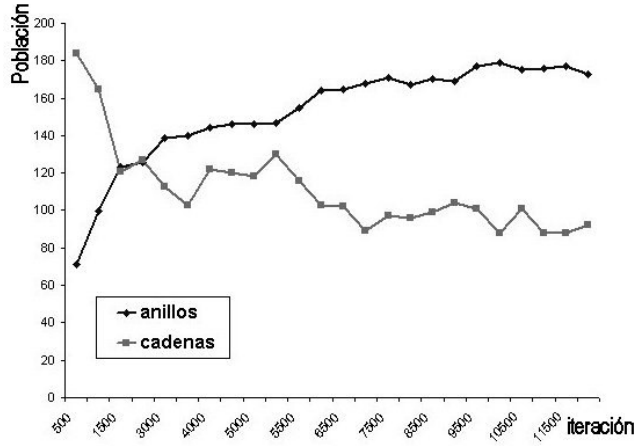


Figura 5.3: Aumento de los anillos a expensas de las cadenas.

5.3. Simulación no espacial

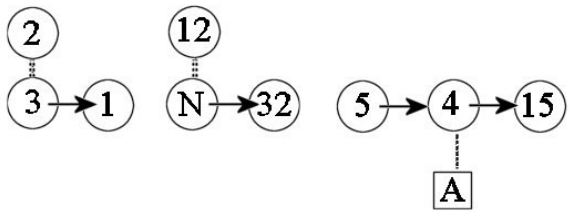
Se pensó en simplificar aún más la simulación espacial, permitiendo superposición de estructuras, eliminando las restricciones de distancia, etc., pero cada una de las simplificaciones reducía la utilidad de tener las partículas insertas en un plano bidimensional. Decidimos por lo tanto pasar a una simulación no espacial, menos realista, pero buscando capturar lo esencial de la dinámica de interés, y permitir correr la simulación a mayor escala. Se usó el lenguaje C++.

#	tipo	L	R	hib	a-á
1	0	3	-	-	-
2	1	-	-	3	-
3	0	-	1	2	-
4	0	5	15	-	1
⋮	⋮	⋮	⋮	⋮	⋮
N	0	-	32	12	-

Cuadro 5.1: Ejemplo de tabla para simulación no espacial.

Para esto se genera una tabla, como la que se muestra en el Cuadro 5.1, en que cada línea corresponde a una base. Para cada base, se señala el tipo, sus enlaces izquierdo y derecho

(si los hay), su base hibridada (si la hay), y el aminoácido asociado (si lo hay). De acuerdo a lo que se ve de la tabla en el Cuadro 5.1, por ejemplo, se puede saber que parte de la situación actual es como lo que se ve en la Figura 5.3 (el resto no lo conocemos, al no estar la tabla completa).



Los aminoácidos no son elementos numerados, sino que aparecen y desaparecen de acuerdo a lo que vaya ocurriendo; se mantiene un registro de la cantidad de aminoácidos "libres" (no asignados), y cuando llega el momento de intentar asociar un aminoácido a una base, éste se crea con un tipo escogido con probabilidad proporcional a la cantidad de aminoácidos "libres" de cada tipo; de este modo se simula un pool de aminoácidos circulando por el entorno. Al deshacerse un codón, el aminoácido simplemente vuelve a sumarse a los libres.

En lugar de encontrarse en un plano bidimensional, aquí las bases que interactúan se escogen al azar desde la tabla. Naturalmente, es necesario introducir un sesgo para favorecer la interacción entre bases que ya están en una misma componente conexa. Por ejemplo, si escojo una base para hibridar, y la tabla indica que su vecino derecho está hibridado a una base que a su vez tiene un vecino derecho, entonces este último es el único candidato para hibridarse con la base escogida (completando el cuadrado). Con criterios de este tipo se simula la presencia de componentes conexas en el plano.

Un problema más delicado es favorecer la formación de ciclos (con N=1.000.000, la probabilidad de que el azar los forme es bajísima). Por lo tanto, habiendo una cadena, hay que crear un sesgo a favor de que el enlace lateral de uno de sus extremos se lleve a cabo con el extremo contrario. Se usaron dos modalidades para eso; en una, se crearon los parámetros *minCycleLength* y *maxCycleLength*, y a la hora de resultar escogida, para enlace lateral, una de las extremidades de una cadena de largo *len*, el ciclo se cerró con probabilidad

$$\frac{len - minCycleLength}{maxCycleLength - minCycleLength}$$

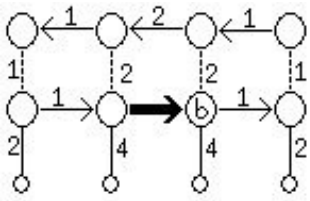
La otra aproximación, un poco más complicada, consistió en suponer que la cadena realiza un "paseo aleatorio" (*random walk*), y que el ciclo se cierra cuando el final de la cadena queda cerca del comienzo. Para ser más preciso, se usó una versión simplificada de los modelos de *random coil* que se usan en el modelamiento de polímeros [Atkins y de Paula, 2002], y que en cada paso aleatorio permiten modificar la dirección sólo dentro de un cierto cono, respecto al paso anterior. Dependiendo de los parámetros, por lo tanto, existe un largo mínimo para los ciclos; sin embargo, no existe un largo máximo.

En general, con cualquiera de los dos métodos, la dependencia de la formación de ciclos respecto de la implementación que se hiciera en el código fue siempre muy fuerte, siendo esta una de las dificultades de la simulación.

Otra dificultad es limitar las estructuras que se pueden generar. Al hacer enlaces laterales e hibridaciones al azar en la tabla puede fácilmente producirse alguna estructura físicamente imposible, o al menos indeseable para nuestros propósitos de imitar un medio bidimensional. Finalmente se optó por rechazar los enlaces (laterales o de hibridación) que generasen estructuras no deseadas, y se programó un test para chequear esto último, según parámetros definidos en el momento de la compilación. Básicamente, las estructuras que se declararon admisibles son las que resultan como subgrafos de un doble anillo (es decir, un anillo, hibridado a su anillo complementario), y también las cadenas simples (sin hibridación con otras cadenas), que tuviesen hibridación consigo mismas a distancia y en orden apropiado para representar hairpins.

A continuación va un esquema simplificado de lo que ocurre en cada iteración de la simulación. Para más detalle se puede consultar el código, disponible con el autor.

- Con probabilidad *probAddAa*, intenta agregar un aminoácido.
  - Escoge una base b al azar.
  - Escoge un tipo de aminoácido a, con probabilidad proporcional a la cantidad de aminoácidos libres de cada tipo.
  - Si b tiene enlaces laterales por ambos costados, y el codón resultante codifica a, se asocia a con b.
- Con probabilidad *probLink*, se intenta crear un enlace lateral.
  - Se escoge una base b_1 al azar.
  - Si no tiene enlace lateral izquierdo, y existe una cadena hacia la derecha, se chequea la condición (descrita arriba) para decidir si se cierra el ciclo.
  - Si no tiene enlace lateral derecho, se hace lo mismo hacia la izquierda.
  - Si no se ha cerrado el ciclo según los criterios anteriores, se escoge una base b_2 al azar, y se unen.
  - Si el cambio ha generado una estructura no aprobada, se revierte.
- Con probabilidad *probUnlink* se intenta romper un enlace lateral.
  - Se escoge una base b al azar.
  - Si b tiene enlace lateral izquierdo, se intentará romper dicho enlace. De lo contrario, se tomará el derecho. Si ninguno existe, no se hace nada.
  - Se calcula el puntaje de supervivencia del enlace, de acuerdo a la estructura circundante (tal como se hacía en la simulación espacial). En este caso el esquema de puntaje incluye la posible cadena hibridada. El esquema que se usó fue el que indica la figura.



(se muestra para el caso de que se quiera romper el enlace lateral izquierdo de b). Cada línea que esté presente suma puntos.

- Con probabilidad proporcional al puntaje obtenido (normalizado por el máximo posible), el enlace se rompe.
- En caso de que al romperlo se deshagan uno o más codones, se liberan los posibles aminoácidos asociados.

- Con probabilidad *probHib* se intenta hibridar bases.
  - Se escoge una base b_1 al azar.
  - Si existe un cuadrado a completar (como se describió antes en el texto), se completa.
  - De lo contrario, y si b_1 es parte de una cadena tipo hairpin (lo que incluye, dada la forma en que se testea, a las cadenas simples), entonces con cierta probabilidad, escoger una base b_2 en la cadena, e hibridar.
  - De lo contrario, escoger una base b_2 al azar, e hibridar.
  - Hibridar *recursivamente* a los vecinos laterales de b_1 y b_2, hasta donde esto sea posible.
  - Si el resultado de los cambios es un estructura no admitida, estos se revierten.
- Con probabilidad *probUnHib* se intenta des-hibridar bases.
  - Se escoge una base b al azar.
  - Se rompen las hibridaciones de *toda* la cadena de la que b forma parte.

Resultados

A diferencia de la simulación espacial, en esta sí fue posible observar algunos resultados, y en general, aunque no son contundentes, apuntan en la dirección señalada en la propuesta de J. Demongeot.

En primer lugar, sí es posible observar en la simulación el efecto de las presiones selectivas sobre los anillos formados. Esto se vio mediante el uso de un código genético de la forma $BBABBABB$ (de modo que sólo los codones 010 y 101 codificaran A), introduciendo solamente aminoácidos de tipo A. El resultado fue una clara preferencia por la formación, en los anillos de largo 4, de anillos de la forma 0101, y en el caso de los anillos de largo

5, una preferencia mayor por el anillo 00101, y una algo menor por 00001[1] (observándose así una correlación con la cantidad de codones del anillo que codifican A), como muestran los gráficos en la Figura 5.4.

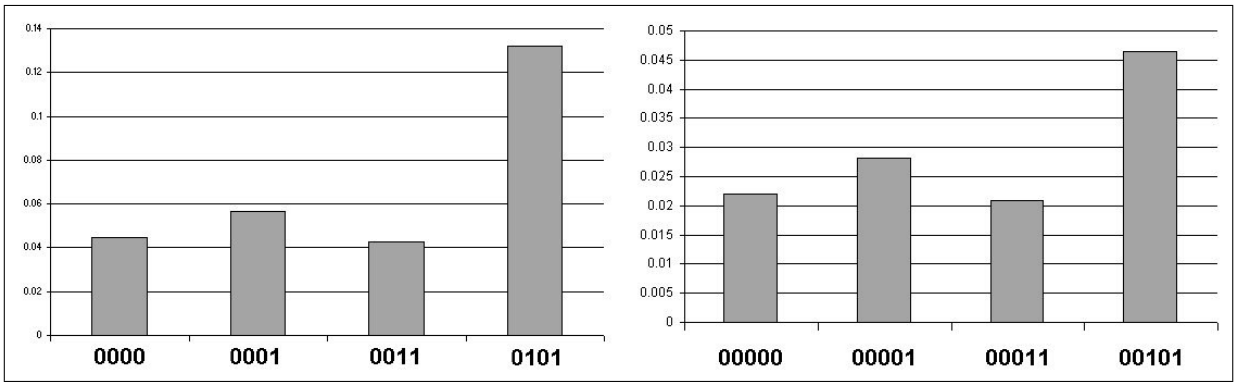


Figura 5.4: Evolución en presencia de un solo aminoácido.

Además, al usar un código genético escogido en $\mathcal{C}_1 \setminus \mathcal{C}_2$, pudo apreciarse la ventaja del anillo 0011 (capaz de codificar un aminoácido de cada tipo), sustentando la idea de J. Demongeot. Sin embargo, la ventaja es leve, como puede verse en el gráfico de la Figura 5.5.

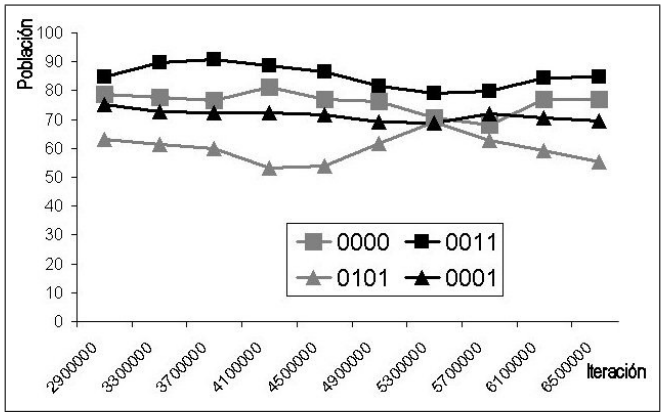


Figura 5.5: Ventaja de codificar un aminoácido por clase.

Finalmente, un resultado interesante es que la producción de anillos 0011 *aumentó* cuando se *eliminó un tipo de aminoácido*, es decir, cuando una de las clases de equivalencia del código no codificó nada. Esto es de interés, pues puede explicar el hecho de que los anillos primitivos hayan optado por codificar, no sólo todos los aminoácidos, sino también la clase de equivalencia formada por los codones de *stop*.

[1]Siempre promediamos los resultados al interior de cada clase de equivalencia, con la relación definida transitivamente por la rotación y la complementación, de modo que, por ejemplo, para anillos de largo cuatro, se consideran cuatro clases, {0000, 1111}, {0001, 0010, 0100, 1000, 0111, 1011, 1101, 1110}, {0101, 1010} y {0011, 0110, 1100, 1001}.

5.4. Direcciones a explorar

Las simulaciones descritas en esta sección son sólo un primer paso, si lo que se pretende es una simulación realista o si se esperan resultados más contundentes.

En primer lugar, sería interesante insistir en una simulación espacial, simplificada con respecto a la nuestra (que al intentar ser realista, pecó de exceso de trabajo computacional). Se podrían eliminar, por ejemplo, algunas restricciones a los movimientos, discretizar las coordenadas, etc.

También podría pensarse en alguna forma de combinación entre la simulación espacial y la no-espacial, reduciendo por ejemplo -en la espacial- los agregados a partículas (y aplicándoles las reglas de auto-interacción de la no espacial), o usando -también en la espacial- un criterio no espacial para las bases libres y los aminoácidos (como se hizo en la no espacial con estos últimos). Por otro lado, podrían introducirse elementos espaciales en la simulación no espacial, usando por ejemplo la distancia de los elementos dentro de la tabla a la hora de seleccionar los participantes en las interacciones aleatorias (de modo de simular el efecto de un medio unidimensional).

También se puede pensar en añadir algunas cosas al modelo; en particular, en el caso de la simulación no espacial, pero también en el de la espacial, sería interesante permitir la formación de péptidos, pues hasta aquí sólo han aparecido aminoácidos aislados. Incluso podría pensarse en incorporar la posibilidad de reproducción por plantilla de estos péptidos, *à la* Harris [Harris et al., 1999], o *à la* Ikehara [Ikehara, 2002b; Ikehara, 2002a]. Incluso sin esto último, sospechamos que la introducción de péptidos aumentaría de manera importante la ventaja de los anillos del tipo de Demongeot por sobre los otros, al introducirse una retroalimentación positiva.

Finalmente, una forma de simulación que no se llevó a cabo pero que podría intentarse, es mediante el listado exhaustivo de las estructuras que pueden aparecer en el modelo, y un listado, exhaustivo también, de sus posibles interacciones y transformaciones, de modo de poder simular (o, en el mejor aunque más improbable de los casos) resolver analíticamente las relaciones entre las proporciones de las distintas estructuras, y determinar el producto a largo plazo a partir de las posibles condiciones iniciales.

Capítulo 6

Búsqueda combinatoria de RNAs minimales

6.1. Anillos minimales bajo distintas condiciones de codificación

En [Weil et al., 1995], J. Demongeot construyó la lista de secuencias de largo 22, iniciadas por **aug** y terminadas con un codón de término, que al cerrarse como anillo pudiesen codificar los 20 aminoácidos. Para hacerlo, y ya que el espacio de búsqueda era de 4^{22} secuencias, se usó un software diseñado para *constraint satisfaction problems*.

Como a nosotros nos interesaba reproducir esto, pero además, chequear la existencia y cantidad de anillos bajo otras restricciones de codificación, optamos por programar una rutina capaz de listar las secuencias de manera directa. En su forma más general, la rutina es llamada con los siguientes argumentos:

```
listaCiclos(
   int length,
   int min[21],
   int max[21],
   int minAnti[21],
   int maxAnti[21],
   char *fixed = 0
)
```

donde `length` es el largo de los anillos buscados, y para cada clase de equivalencia (sinónima) del código genético, es decir, para los 20 aminoácidos y el *stop*, `min[aa]` y `max[aa]` indican la cantidad mínima y máxima de veces que la clase puede ser codificada por el anillo. Los arreglos `minAnti` y `maxAnti` cumplen la misma función pero imponen la restricción sobre el anillo *complementario*, y finalmente, `fixed` es una cadena optativa que

se puede imponer como parte del anillo. Al escribir las llamadas a la rutina, usaremos la notación $\vec{\alpha} = \{\alpha, \ldots, \alpha\}$.

Para poder buscar en las $4^{\texttt{length}}$ secuencias posibles, se recorre el árbol de secuencias posibles, cortando las ramas que se vayan probando inservibles (cuando la cantidad codificada para alguna clase exceda su cota superior, o cuando el largo de la secuencia ya fijada permita ver que será imposible satisfacer las cotas inferiores). De este modo fue posible encontrar las soluciones incluso para condiciones bajo las cuales los anillos minimales eran de largo 26 (no hizo falta, bajo ninguna de las condiciones, intentar anillos aún mayores), aunque en un caso de largo 26 el tiempo de ejecución fue de casi dos días.

6.1.1. A_1: Los 1280 anillos de Demongeot

Primero que nada, se utilizó esta rutina para volver a generar los anillos buscados originalmente por Demongeot, separando los anillos según el codón de término, con las siguientes llamadas a la rutina:

$$\begin{aligned} A_{uaa} &= \texttt{listaCiclos}(22, \{1, \ldots, 2, \ldots, 1\}, \vec{\infty}, \vec{0}, \vec{\infty}, \text{“uaaaug”}) \\ A_{uag} &= \texttt{listaCiclos}(22, \{1, \ldots, 2, \ldots, 1\}, \vec{\infty}, \vec{0}, \vec{\infty}, \text{“uagaug”}) \\ A_{uga} &= \texttt{listaCiclos}(22, \{1, \ldots, 2, \ldots, 1\}, \vec{\infty}, \vec{0}, \vec{\infty}, \text{“ugaaug”}) \end{aligned}$$

con el ‘$2'$ de los mínimos puesto en metionina (M), que debe usarse de *start* y además codificarse. La generación de los tres listados tomó, en conjunto, menos de 2 segundos con un procesador a 2 GHz; bastante menos que los 9433 que tomó en [Weil et al., 1995], aún considerando la diferencia atribuible al avance en los procesadores desde ese año; al parecer, este tipo de listado directo es eficiente.

Los conjuntos obtenidos, por supuesto, son los mismos que en [Weil et al., 1995], con $\#A_{uaa} = 256$, $\#A_{uag} = 352$ y $\#A_{uga} = 672$, haciendo un total de 1280; llamaremos A_1 a la unión de los tres. Se pueden hacer de inmediato algunas observaciones sobre el contenido de estos conjuntos.

Conjunto	a	c	g	u
A_{uaa}	8	3	5	6
A_{uag}	8	3	5	6
A_{uga}	7	4	5	6
Promedio A_1	7.5	3.5	5	6

Cuadro 6.1: *Uso de bases en los anillos de Demongeot.*

En primer lugar, en cada uno de ellos, todas las secuencias comparten un mismo uso de las bases (ver Cuadro 6.1). De hecho, comparten un mismo uso de las palabras de largo 2 (“dinucleótidos”), indicado en el Cuadro 6.2, y éste es incluso idéntico en A_{uaa} y A_{uag}; esto es una muestra de lo restrictiva que resulta la combinatoria, con la exigencia de recorrer todas las clases de equivalencia del código.

Conjunto	aa	ac	ag	au	ca	cc	cg	cu	ga	gc	gg	gu	ua	uc	ug	uu
A_{uaa}	2	1	2	3	2	1	0	0	2	1	1	1	2	0	2	2
A_{uag}	2	1	2	3	2	1	0	0	2	1	1	1	2	0	2	2
A_{uga}	2	1	1	3	2	1	0	1	2	1	1	1	1	1	3	1
Promedio A_1	2	1	1.5	3	2	1	0	0.5	2	1	1	1	1.5	0.5	2.5	1.5

Cuadro 6.2: Uso de dinucleótidos en los anillos de Demongeot.

Los codones usados, en cambio, no son los mismos al interior de cada conjunto. En A_{uaa}, por ejemplo, 11 aminoácidos son codificados siempre por el mismo codón, pero para los 9 restantes existen 2 alternativas, que se usan con las frecuencias indicadas en el Cuadro 6.3.

Aminoácido	Codón	# anillos	Frecuencia
T	aca	128	0.5
	acc	128	0.5
S	agc	96	0.375
	agu	160	0.625
I	aua	128	0.5
	auu	128	0.5
H	cac	48	0.1875
	cau	208	0.8125
D	gac	96	0.375
	gau	160	0.625
A	gca	128	0.5
	gcc	128	0.5
V	gua	128	0.5
	guu	128	0.5
Y	cac	112	0.4375
	cau	144	0.5625
C	ugc	160	0.625
	ugu	96	0.375

Cuadro 6.3: Uso de codones en A_{uaa}.

Las cantidades de uso de los codones en A_{uaa}, con potencias de dos y múltiplos pequeños de éstas, sugieren la presencia de una combinatoria bastante simple, y en efecto, ésta existe. De las $2^9 = 512$ posibles combinaciones de uso de estos codones, sólo 24 ocurren, y se generan por tres pares de codones que toman sólo dos posibilidades cada uno (el codón para T termina en **c** cuando el de A termina en **a**, y viceversa; S usa **c** cuando C usa **u** y viceversa; I usa **a** cuando V usa **u** y viceversa), y un trío, H, D e Y, donde de los tres codones, siempre dos terminan en **u** y el otro en **c**. El resto de la combinatoria consiste en cambios de orden de estas piezas. Para A_{uag} y A_{uga} existen esquemas parecidos, aunque un poco más complicados. En general, los cambios de orden de los codones son muchos; la mayoría de los aminoácidos puede ubicarse en casi todas las posiciones del anillo, como muestra el Cuadro 6.4 para A_{uaa}, y por supuesto, esto se ve reflejado en un variado uso de las bases a lo largo de los anillos. En la Figura 6.1 puede verse la frecuencia de uso de

las distintas bases en los anillos de A_1 (salvo por las posiciones fijas de **aug** y el codón de término).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
A	0	80	0	0	32	12	1	17	20	18	17	11	16	14	10	8	0	0
C	112	0	0	0	16	0	20	8	10	11	9	8	22	5	19	0	16	0
D	0	0	48	0	56	8	7	21	18	9	13	9	8	5	30	0	24	0
E	0	0	96	0	4	4	19	17	14	9	7	27	10	23	6	20	0	0
F	0	0	16	0	9	13	12	5	10	12	9	6	12	0	24	0	128	0
G	0	144	0	0	0	18	10	13	6	5	14	9	10	13	8	6	0	0
H	0	0	24	24	5	13	18	19	12	12	38	7	13	1	38	0	32	0
I	0	0	0	18	6	16	6	8	16	6	4	16	0	0	0	80	0	80
M	0	0	0	34	10	33	14	15	25	18	18	35	13	25	0	16	0	0
L	0	0	0	16	0	9	13	12	5	10	12	9	6	12	0	24	0	128
P	0	0	40	0	22	16	27	8	21	23	21	16	15	25	6	16	0	0
Q	0	0	16	16	9	33	20	15	15	23	10	31	18	26	12	12	0	0
R	0	0	0	60	12	8	28	19	12	15	22	9	18	23	12	18	0	0
S	0	0	0	52	4	5	9	20	20	14	10	8	40	5	37	0	32	0
T	0	0	0	36	8	37	14	23	17	28	16	19	21	17	12	8	0	0
V	0	32	0	0	20	8	4	12	8	12	8	8	0	48	0	48	0	48
W	144	0	0	0	18	10	13	6	5	14	9	10	13	8	6	0	0	0
Y	0	0	16	0	25	13	21	18	22	17	19	18	21	6	36	0	24	0

Cuadro 6.4: Ubicación de los aminoácidos en A_{uaa}.

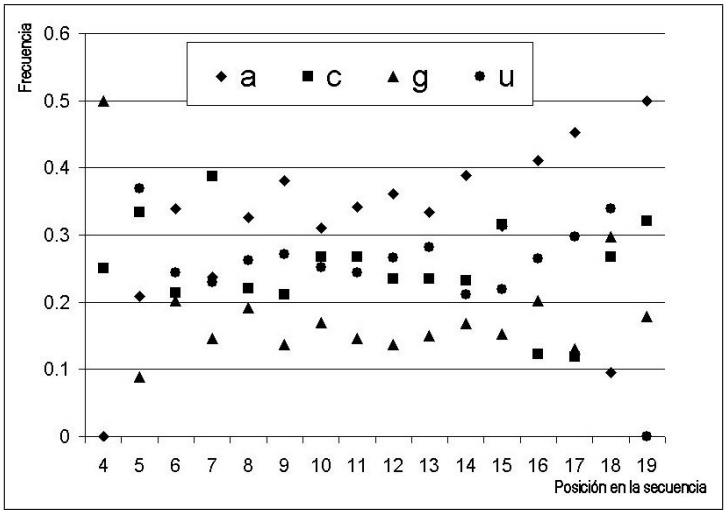


Figura 6.1: Frecuencia de las bases a lo largo de A_1 (abriendo cada anillo de modo que el codón de término queda al final).

Otro cálculo interesante es la cantidad de aminoácidos codificados por el anillo *complementario* a cada uno de los anillos listados. Después de todo, en el escenario planteado por Demongeot no sólo deben formarse los anillos, sino que deben reproducirse, y por lo tanto, es de esperar que la secuencia complementaria tenga también "buenas propiedades". Al listar los anillos complementarios, se encuentran 140 que codifican 18 clases de equivalencia (aminoácidos o stop), 760 que codifican 17, 304 que codifican 16, 64 que

codifican 15, y sólo 12 que codifican 14; ninguno codifica menos. Para ver si esto distingue a A_1 de una población de anillos aleatorios, estimamos (con una población de 3×10^7 anillos generados al azar) la probabilidad de que se codifiquen k clases de equivalencia distintas, con $k = 0, \ldots, 22$. Para incluir el efecto de la cantidad de bases de cada tipo empleadas, estimamos también la distribución de probabilidades usando una población de 100.000 anillos aleatorios con el mismo uso de bases que A_{uaa} y A_{uag} (es decir, 8 veces **a**, 3 veces **c**, 5 veces **g**, 6 veces **u**); el resultado está en la Figura 6.2. Se nota claramente cómo la restricción impuesta a A_1 tiene también como consecuencia sesgar fuertemente la distribución, contribuyendo a la codificación de muchos aminoácidos en el anillo complementario, lo que calza con la hipótesis de Demongeot. Finalmente en la Figura 6.3 podemos ver cuáles son las clases de equivalencia que se están codificando en estos anillos complementarios (para D, la frecuencia es 0) .

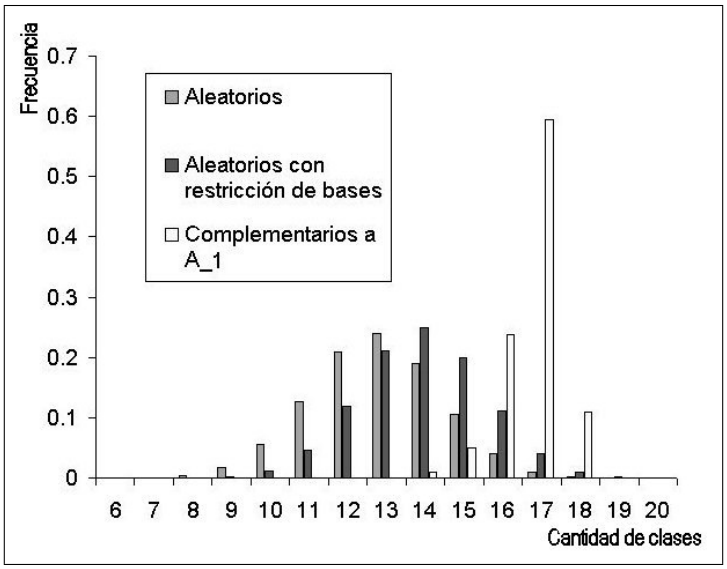


Figura 6.2: Frecuencia con que se codifican distintas cantidades de clases de equivalencia del código genético, en anillos aleatorios y en los anillos complementarios a A_1.

6.1.2. Otros restricciones de codificación

No existen anillos de largo 20 que codifiquen simultáneamente los 20 aminoácidos. Esto, que ya se había notado en [Weil et al., 1995], fue confirmado al obtenerse una lista vacía como resultado de la siguiente llamada, poniendo $k = 20$:

$$\texttt{listaCiclos}(k, \{1, \ldots, 1, 0\}, \vec{\infty}, \vec{0}, \vec{\infty})$$

Por otro lado, sí existen soluciones (3456) cuando se pone $k = 21$, y *todas* ellas repiten el aminoácido T, triptofano. La codificación de clases distintas en los anillos complementarios resulta similar a la de A_1, aunque incluso más cargada hacia un número alto de clases:

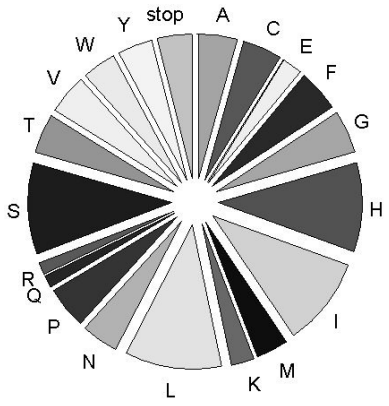


Figura 6.3: Frecuencia de ocurrencia de las clases de equivalencia del código, en los anillos complementarios a A_1.

128, 452, 728, 1164 y 984 anillos complementarios codifican 14, 15, 16, 17 y 18 clases distintas, respectivamente.

Al llamar a la rutina como arriba, con $k = 22$, el resultado son 196848 anillos (que obviamente incluyen a A_1). Triptofano sigue siendo el aminoácido más repetido, aunque también otros aparecen con frecuencia (ver Figura 6.4).

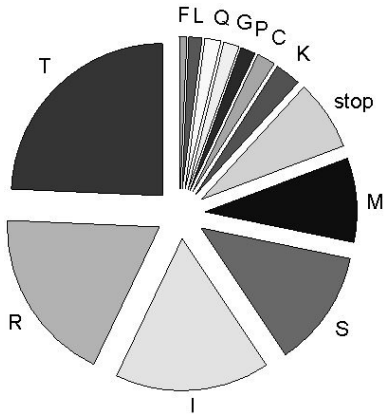


Figura 6.4: Frecuencia de repetición de las clases de equivalencia del código, al codificar 20 aminoácidos con anillos de largo 22.

Si queremos considerar el tema de la codificación de aminoácidos en los anillos complementarios, una pregunta legítima es respecto a los anillos minimales que puedan codificar los 20 aminoácidos, *en ambas direcciones*. En otras palabras, pedir el primer k para el cual la siguiente llamada entrega soluciones:

$$\texttt{listaCiclos}(k, \{1, \ldots, 1, 0\}, \vec{\infty}, \{1, \ldots, 1, 0\}, \vec{\infty})$$

Dicho k resulta ser 24, y el conjunto contiene 69280 anillos. En realidad, sólo son 34640

anillos distintos, pues para cada anillo, su anillo complementario está también en el conjunto, y hemos verificado que ninguno es idéntico a su anillo complementario (la mínima distancia de Hamming[1] entre estos y sus complementos es 4, alcanzada por el 1.4 % de los anillos). Los aminoácidos I y S se repiten en todos estos anillos; las otras repeticiones son F, L, R y nuevamente S (en porcentajes de 14.8, 35.2, 30.5 y 19,5, respectivamente).

También se buscó el menor k que entregara soluciones con la llamada

$$\texttt{listaCiclos}(k, \vec{1}, \vec{\infty}, \vec{1}, \vec{\infty}),$$

es decir, anillos que codifiquen las 21 clases de equivalencia del código en ambas direcciones. En este caso el menor k es 26, y se obtienen 525536 soluciones; los aminoácidos que se repiten son L, S, I, R, H, F y M, con 31, 30, 22, 5, 5, 4 y 3 por ciento respectivamente.

Otro listado que se elaboró, usando una versión modificada de `listaCiclos`, fue de anillos minimales que *no usaran* ciertos codones. Esto nace de una pregunta propuesta por J. Demongeot, preguntándose acaso los codones no ocupados por el anillo AB (su elección dentro de A_1) podían formar anillos más pequeños, sub-optimales, que fuesen también parte de la población original. Si existiese, por ejemplo, un anillo de largo 19 codificando 18 aminoácidos y un stop (pues para dos aminoácidos, los únicos codones posibles ya se usaron en AB), y luego otro de largo menor, al que se le prohibiesen los codones de AB y los del anillo de 19, y codificase todas las clases de equivalencia aún disponibles, y así sucesivamente, podría pensarse en una estructura de anillos que reflejase de alguna manera (y al vez ayudase a explicar) la degeneración del código genético. (Es por esto que también en las consideraciones sobre códigos binarios -ver 5.1- vimos los casos en que podían darse semejantes anillos sub-optimales.) Al hacer este cómputo encontramos que los anillos más pequeños capaces de codificar 18 aminoácidos y un stop se obtienen para largo 23 y el stop **uaa**, y son 72. Si se pide que el stop sea **uag**, entonces el largo mínimo es 26, y se obtienen 928 soluciones. Eso deja en mal pie la idea de una serie de anillos, progresivamente más pequeños.

Otro cómputo que se hizo, con anillos más pequeños, fue buscar anillos en que todos los codones codificasen cosas distintas. Ya sabemos que con 20 letras no es posible codificar los 20 aminoácidos. ¿Es posible codificar, por ejemplo, 19 de ellos, con 19 letras? En principio puede parecer igual de dificil vencer las restricciones de la combinatoria; sin embargo, lo que se encuentra es lo contrario. Concretamente, lo que se construyó fueron los siguientes conjuntos:

$$\begin{aligned}
\mathrm{A}^k &= \texttt{listaCiclos}(k, \vec{0}, \{1, \ldots, 1, 0\}, \vec{0}, \vec{\infty}) \\
\mathrm{A}^k_* &= \texttt{listaCiclos}(k, \{0, \ldots, 0, 0\}, \vec{1}, \vec{0}, \vec{\infty}) \\
\mathrm{A}^k_x &= \texttt{listaCiclos}(k, \{0, \ldots, 0, 1(\mathrm{M}), 0, \ldots, 0\}, \vec{1}, \vec{0}, \vec{\infty}, \text{“[stop]aug”}) \\
\mathrm{A}^k_+ &= \texttt{listaCiclos}(k, \{0, \ldots, 0, 2(\mathrm{M}), 0, \ldots, 0\}, \{1, \ldots, 1, 2(\mathrm{M}), 1, \ldots, 1\}, \vec{0}, \vec{\infty}, \text{“[stop]aug”})
\end{aligned}$$

En otras palabras: A^k son los anillos de largo k que codifican k aminoácidos diferentes. A^k_* son los anillos de largo k que codifican k clases de equivalencia distintas (es decir, los

[1] Ver 6.2.1.

aminoácidos y stop), y obviamente, $\mathrm{A}^k \subseteq \mathrm{A}^k_*$. A^k_x son anillos de largo k y de la forma "**aug**... stop" (como los de A_1) que codifican $k-1$ aminoácidos distintos, y A^k_+ son los anillos de esa forma, que codifican $k-2$ aminoácidos distintos, permitiendo una repetición de M (es decir, dejándolo una vez sólo como señal de inicio), de modo que $\mathrm{A}^{22}_+ = \mathrm{A}_1$. El tamaño de los conjuntos resultantes se presenta en el Cuadro 6.5. En él se puede ver lo curiosos que resultan los conjuntos vacíos en los largos 20 y 21; la causa está sólo en la combinatoria resultante del código genético.

k	A^k	A^k_*	A^k_x	A^k_+
3	13	16	-	-
4	35	47	-	-
5	81	120	-	-
6	193	306	3	0
7	428	732	9	0
8	889	1658	22	3
9	1724	3578	59	7
10	3110	7166	154	23
11	5248	13315	336	64
12	8205	22918	701	189
13	11708	36059	1319	453
14	14647	51020	2140	1013
15	16128	64416	2988	1989
16	15756	70418	3336	3479
17	12644	62928	3144	5440
18	6664	43706	2266	6812
19	1488	17984	1054	8918
20	0	3264	216	8324
21	0	0	0	0
22	-	-	-	1280

Cuadro 6.5: Cantidad de anillos pequeños bajo distintas restricciones de codificación.

6.1.3. A_0: Generalizando los anillos de Demongeot

El último conjunto de anillos generado, y que resultó ser de sumo interés, fue

$$\mathrm{A}_0 \;=\; \texttt{listaCiclos}(22, \vec{1}, \vec{\infty}, \vec{0}, \vec{\infty})$$

Es decir, pedir que se codifiquen las 21 clases de equivalencia del código. Esto es imposible para anillos de largo 21 (como se ve en el Cuadro 6.5), pero sí es posible en largo 22 (como atestiguaba ya A_1, que está incluído en A_0). Lo que quisimos fue ver qué pasaba si *no imponíamos* la restricción, impuesta para A_1, de que los anillos pudiesen escribirse en la

forma "**aug**... stop", pero sí exigíamos a los anillos codificar stop, además de los aminoácidos. Las razones para pensar en esto fueron dos: por un lado, la exigencia de Demongeot de pedir anillos con la forma recién dicha resulta extremadamente arbitraria, dado que en el escenario que se plantea, aún no existe ninguna maquinaria celular para la cual esa forma pueda significar algo. La segunda razón fue la observación hecha en simulaciones (5.3) de que la existencia de una clase que *no tuviese* un aminoácido asociado (como es la del stop) puede servir de incentivo a la reproducción de los anillos (probablemente al permitir un punto de partida para la replicación). Por lo tanto, una exigencia razonable para los anillos primigenios podía ser la codificación de las 21 clases de equivalencia.

El resultado, A_0, contiene 29520 anillos. De estos, 15120 usan el codón **uga** para codificar stop, mientras que 7584 usan **uaa** y 6816 usan **uag**. Para los anillos que usan **uga**, el uso de dinucleótidos (y de bases, por lo tanto) es siempre el mismo, coincidiendo con el señalado para A_{uga} en los Cuadros 6.1 y 6.2. Lo mismo pasa con los restantes anillos: su uso de dinucleótidos es siempre el mismo, coincidiendo con el tabulado para A_{uaa} y A_{uag} en dichos cuadros.

Al igual que para A_1, calculamos la cantidad de clases codificadas por los anillos complementarios; el resultado son 144 anillos codificando 14 clases, 1392 codificando 15, 7608 codificando 16, 16416 codificando 17, y 3960 codificando 18. Con esto, el promedio de clases codificadas por los anillos complementarios resulta incluso un poco superior al valor para A_1. Casi todas las clases (salvo D) aparecen en los anillos complementarios, en la proporción indicada en la Figura 6.5.

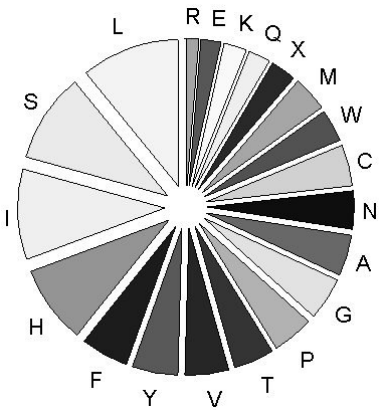


Figura 6.5: Frecuencia de ocurrencia de las clases de equivalencia del código, en los anillos complementarios a A_0.

Sin embargo, más interesante es lo que ocurre con el aminoácido que debe repetirse *al derecho*: siempre corresponde a un codón de la forma **auN**, y en el *52 %* de los casos, este codón es **aug**, *es decir, el codón de inicio*. Esto sugiere que tal vez ese codón resultó ser el de inicio por haber estado "disponible" (esto es, repetido) en los anillos primitivos, y sería una posible explicación (y primera, hasta donde llega nuestro conocimiento) para esta arbitrariedad del código. Desde aquí será A_0, y no ya A_1, el conjunto de anillos de

mayor interés.

6.2. Hairpins y palindromía

En el escenario planteado por J. Demongeot, existen cadenas y anillos de RNA que están permanentemente formándose, rompiéndose, hibridándose, etc. En ese escenario, no sólo es importante poder estabilizarse mediante la afinidad con aminoácidos, sino también, tener una probabilidad mayor de formación en virtud de la frecuencia con que otras cadenas aparezcan en el medio (tal vez por destrucción previa de anillos del mismo tipo). Es por eso que sería "útil" para un anillo, por ejemplo, ser muy parecido a su propio complemento (pues por hibridación, también el complemento estará presente en el medio), es decir, ser un *palíndrome*, no en el sentido de la reversión directa de la secuencia, sino en el sentido de la complementaridad (reversión sumada al pareo de Watson-Crick).

Esto nos lleva a preguntarnos por la similitud existente entre las secuencias y sus complementos, y en general, a la existente también entre unas secuencias y otras. Por ello, la primera sección discutirá algunas posibles medidas de distancia entre anillos (métricas o semi-métricas), que he diseñado con el propósito de capturar los diversos aspectos del juego combinatorio del escenario primitivo. Es interesante notar que, sin quererlo, el resultado fueron tres funciones que de alguna forma se corresponden con los principales mecanismos de evolución de los genomas: mutación (d_H^c, d_H), transposición (d_s) y rebarajamiento (*shuffling*) (d_t, $\hat{d}_t$). Una de ellas, la computacionalmente más liviana, se evaluará a lo largo del resto de esta sección, pero es en 6.3 donde resultarán decisivas.

Por otro lado, también debió ser útil para las secuencias ser capaces de "encapsularse" en una forma químicamente estable, ante la posibilidad de tiempos difíciles (cambios en la temperatura u otras condiciones del ambiente). Esto pudo lograrse mediante la formación de hairpins, la forma más estable para un RNA pequeño. Por lo tanto, interesa conocer la capacidad de los anillos de A_1 ó A_0 para formar buenos hairpins.

6.2.1. Medición de distancias entre anillos

Nos interesa buscar formas de comparar anillos; es decir, nos interesa definir nociones de distancia entre vectores, pero donde dos vectores son equivalentes si uno es una rotación del otro. Trabajaremos con un alfabeto finito A fijo, y con un entero n fijo también, que denotará la longitud de los anillos. Los vectores, por lo tanto, estarán en A^n.

Definamos primero la notación para la rotación: dado $x \in A^n$, definimos

$$\sigma(x) \;=\; (x_1, \ldots, x_{n-1}, x_0)$$

Es trivial ver que σ es invertible, que $\sigma^i(\sigma^j(x)) = \sigma^{i+j}(x)$, y que $\sigma^i(x) = \sigma^{i \bmod n}(x)$.

Definiremos la noción de equivalencia bajo rotación, notada por "$\equiv$", para dos vectores $x, y \in A^n$, por

$$x \equiv y \iff \exists k : x = \sigma^k(y)$$

Es trivial ver que esta es en efecto una relación de equivalencia. Nuestro espacio de anillos, entonces, será $(A^n, \equiv)$, el conjunto cuociente formado por las clases de equivalencia de los vectores, y un anillo será escrito como $[x]$ (aunque fuera de esta sección, usaremos x indistintantemente para designar una cadena o el anillo que representa –pues este último uso es con mucho el más frecuente en el resto del texto).

Distancia de Hamming circular

La forma usual de comparar vectores que toman sus valores en un alfabeto finito es mediante la *distancia de Hamming*. Dados dos vectores $x, y \in A^n$, la distancia de Hamming entre ellos es

$$d_H(x, y) \;=\; \#\left\{i \in \{0, \dots, n-1\} \;:\; x_i \neq y_i\right\}$$

Es decir, es la cantidad de posiciones en que los valores de los vectores difieren. La función d_H es una métrica: es no-negativa, simétrica, satisface la desigualdad triangular, y una distancia nula implica la igualdad entre los vectores. Es fácil ver que $d_H(x, y) = d_H(\sigma^i(x), \sigma^i(y))$, y que por lo tanto $d_H(x, \sigma^i(y)) = d_H(\sigma^{-i}(x), y)$. Usando esto, definiremos la *distancia de Hamming circular* entre dos anillos $[x], [y]$ como

$$d_H^c([x], [y]) \;=\; \min_k d_H(x, \sigma^k(y)) \;=\; \min_{0 \leq k \leq n-1} d_H(x, \sigma^k(y))$$

En general, el mínimo entre dos o más métricas no es necesariamente una métrica (como se muestra en la Figura 6.6). Por lo tanto, debemos mostrar que d_H^c efectivamente lo es.

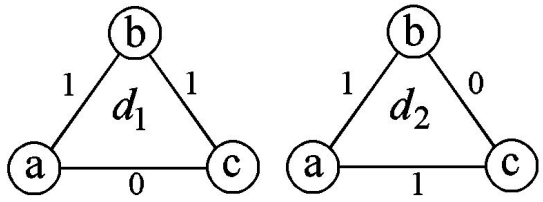


Figura 6.6: El mínimo entre d_1 y d_2 no satisface la desigualdad triangular: $1 = \min\{d_1(a,b), d_2(a,b)\} > \min\{d_1(a,c), d_2(a,c)\} + \min\{d_1(c,b), d_2(c,b)\} = 0$.

Lema 1: d_H^c es una métrica en $(A^n, \equiv)$.

Demostración.

1. Si $d_H^c([x], [y]) = 0$, esto implica que existe k tal que $d_H(x, \sigma^k(y)) = 0$; por lo tanto, $x = \sigma^k(y)$ y $[x] = [y]$.

2. Simetría:

$$d_H^c([x],[y]) \;=\; \min_k d_H(x,\sigma^k(y)) \;=\; \min_k d_H(\sigma^{-k}(x),y) \;=\; \min_k d_H(y,\sigma^{-k}(x)) \;=\; d_H^c([y],[x])$$

3. Sean $[x],[y],[z] \in (A^n,\equiv)$. Debemos demostrar la desigualdad triangular, es decir, que $d_H^c([x],[y]) \leq d_H^c([x],[z]) + d_H^c([z],[y])$. Sean i, j tales que

$$d_H^c([z],[x]) \;=\; d_H(z,\sigma^i(x)) \quad , \quad d_H^c([z],[y]) \;=\; d_H(z,\sigma^j(y))$$

Definamos además

$$\begin{array}{l} a = \#\{k : \sigma^i(x)_k \;\neq\; \sigma^j(y)_k \;=\; z_k\} \\ b = \#\{k : \sigma^i(x)_k \;=\; \sigma^j(y)_k \;\neq\; z_k\} \\ c = \#\{k : z_k \;=\; \sigma^i(x)_k \;\neq\; \sigma^j(y)_k\} \\ d = \#\{k : \sigma^i(x)_k \;\neq\; z_k,\; \sigma^j(y)_k \;\neq\; z_k,\; \sigma^i(x)_k \;\neq\; \sigma^j(y)_k\} \end{array}$$

Entonces

$$d_H(\sigma^i(x),\sigma^j(y)) \;=\; a{+}c{+}d \;\leq\; (a{+}b{+}d){+}(b{+}c{+}d) \;=\; d_H(z,\sigma^i(x)){+}d_H(z,\sigma^j(y))$$

y por lo tanto

$$d_H^c([x],[y]) \;\leq\; d_H(\sigma^i(x),\sigma^j(y)) \;=\; d_H(z,\sigma^i(x)){+}d_H(z,\sigma^j(y)) \;=\; d_H^c([z],[x]){+}d_H^c([z],[y])$$

□

Distancia por subcadenas maximales

Definiremos también otra medida de distancia, que denotaremos d_s, y evalúa la existencia de una subcadena común a los dos anillos; más precisamente, definimos $d_s([x],[y])$ como n menos la longitud de la mayor cadena compartida por los anillos $[x]$ e $[y]$, es decir,

$$d_s([x],[y]) \;=\; n - \max_{i,j}\{m \in \{0,\ldots,n\} \;:\; \sigma^i(x)_k = \sigma^j(x)_k \text{ para } 0 \leq k < m\}$$

Es fácil ver que *la función d_s es una semi-métrica en* $(A^n,\equiv)$. No es una métrica, pues la desigualdad triangular puede fallar cuando las subcadenas compartidas por $[z]$ con $[x]$ e $[y]$ tienen como intersección dos subcadenas disconexas (es decir, en conjunto las subcadenas compartidas recubren z, y se intersectan por ambos extremos; un ejemplo se muestra en el cuadro que sigue). Sin embargo, en nuestras evaluaciones sobre una tabla de 34×34 con anillos de largo 22, d_s verificó la desigualdad en el 99.9 % de los casos, lo que sugiere que es una semi-métrica con un comportamiento muy similar al de una métrica.

La desigualdad triangular puede fallar	$[x] = [\texttt{abca}]$	$d_s([x],[y]) = 3$
para d_s : con estos valores,	$[y] = [\texttt{cdac}]$	$d_s([x],[z]) = 1$
$d_s([x],[y]) > d_s([x],[z]) + d_s([z],[y])$	$[z] = [\texttt{abcd}]$	$d_s([z],[y]) = 1$

Lema 2: $d_H^c \leq d_s$

Demostración. Ya que $d_H(x,y) = \#\{i : x_i \neq y_i\}$, también se tiene $n - d_H(x,y) = \#\{i : x_i = y_i\}$, y por lo tanto d_H^c se puede escribir

$$d_H^c([x],[y]) \;=\; \min_k n - \#\{i : x_i = \sigma^k(y)_i\} \;=\; n - \max_k \#\{i : x_i = \sigma^k(y)_i\}$$

Si la mayor cadena que comparten $[x]$ e $[y]$ es de largo m, entonces es fácil ver que

$$\max_k \#\{i :\; x_i = \sigma^k(y)_i\} \;\geq\; m$$

y por lo tanto

$$d_H^c([x],[y]) \;=\; n - \max_k \#\{i : x_i = \sigma^k(y)_i\} \;\leq\; n - m \;=\; d_s([x],[y])$$

□

Distancia por reordenamiento de cadenas

Hasta aquí, las dos funciones presentadas miden de alguna forma semejanza entre los anillos, pero cada una tiene tanto ventajas como desventajas. La distancia de Hamming circular d_H^c mide la similitud entre los anillos, pero pasa por alto el orden de las distintas posiciones; podríamos aplicar una misma permutación a ambos anillos y la distancia se mantiene idéntica. Por lo tanto, en un escenario en que las secuencias se cortan en trozos que se mezclan y luego vuelven a unirse en nuevos anillos, d_H^c no captura mucho de lo que pasa con esas subcadenas de los anillos (sólo en la medida de que admitamos cierta fracción de hibridaciones erróneas, que son posibles pero inestables). Por otro lado, d_s mide el tamaño de la mayor subcadena compartida por dos anillos, pero no nos dice nada sobre el resto de sus secuencias.

Por lo tanto, para poder capturar otro aspecto de la idea de semejanza que nos interesa, introduciremos (informalmente, aunque la idea es clara) una tercera función, d_t. Esta será finita sólo para pares de anillos $[x],[y]$ *que usen la misma cantidad de letras (de cada tipo)*, es decir, tales que $d_H(x,\vec{\alpha}) = d_H(y,\vec{\alpha})$ para todo α en A; valdrá ∞ en caso contrario. Cuando estemos en el caso finito, definiremos $d_t([x],[y])$ como la *cantidad mínima de cortes* que se deben hacer en $[x]$ para que, reordenando los trozos, se obtenga $[y]$.

Lema 3: d_t **es una métrica en** $(A^n, \equiv)$

Demostración (bosquejo). Si $d_s([x],[y]) = 0$, es porque ningún corte es necesario, y por lo tanto los anillos son idénticos. La simetría es fácil de ver, pues los trozos que se usan al ir en cualquiera de las dos direcciones son los mismos. Finalmente, para ver la desigualdad triangular, $d_t([x],[y]) \leq d_t([x],[z]) + d_t([z],[y])$, basta tomar $[x]$, trozarlo de la manera óptima para pasar a $[z]$, y luego hacer *además* los cortes que permiten pasar de manera óptima de $[z]$ a $[y]$. Con esto, se pasa de $[x]$ a $[y]$ con $d_t([x],[z]) + d_t([z],[y])$ cortes; puede no ser la manera óptima de pasar de $[x]$ a $[y]$, pero acota superiormente a

$d_t([x],[y])$, y prueba la desigualdad. Lo anterior es para el caso en que todo es finito; si el lado izquierdo de la desigualdad es infinito, entonces el uso de las letras es distinto en $[x]$ e $[y]$, y como no podrían compartir ambos el mismo uso de letras que $[z]$, el lado derecho tendrá que ser infinito también. □

La semi-métrica $\hat{d}_t$: Para efectos prácticos, a fin de tener una función relativamente eficiente en términos computacionales, lo que efectivamente evaluamos fue un algoritmo que captura sólo parcialmente la idea de d_t, y que describimos a continuación. Dados dos anillos $[x]$ e $[y]$, eliminamos de cada uno la mayor subcadena que compartan, dejando dos palabras x^1, y^1. Con ellas inicializamos dos listas de palabras, $P_x = \{x^1\}$, $P_y = \{y^1\}$. En cada momento, estas listas contendrán una familia de subcadenas disjuntas de $[x]$ e $[y]$, respectivamente. En cada iteración, el algoritmo busca la mayor subcadena común entre dos palabras, tomadas una de cada lista (se maximiza sobre todos los pares de palabras), elimina esa subcadena de las respectivas palabras, y devuelve los trozos sobrantes a las listas respectivas. Definimos la función $\hat{d}_t$ como la *cantidad de iteraciones* del algoritmo hasta que logra vaciar las listas[2].

Es fácil entender por qué llamamos a esta función $\hat{d}_t$: representa la misma idea de d_t, cortar las secuencias en las partes que hagan falta para poder rearmar una a partir de la otra. Y si bien $\hat{d}_t$ es sólo una semi-métrica (por fallar en la desigualdad triangular) verificó dicha desigualdad en el 99.7 % de los triángulos de una tabla de 34×34 con la que trabajamos.

6.2.2. Distancias de A_1 y A_0 respecto a sus complementos y AB

Calculamos, para cada anillo $x \in A_0$ (que incluye a A_1), la distancia d_H^c respecto a su complemento. El resultado fueron valores que van de 4 a 14, alcanzados en proporciones similares en cada conjunto (ver Cuadro 6.6). Para comparar con lo que ocurre en secuencias aleatorias, se generó un listado al azar con 60000 anillos (AZ); para agregar el sesgo provocado por la cantidad de bases de cada tipo usadas, se generó también un listado al azar de 60.000 anillos con el uso de bases de A_{uaa} ("AZ_b"). Los resultados están graficados en la Figura 6.7.

Se puede observar que los anillos de nuestros conjuntos tienen un sesgo hacia una mayor palindromía que la de los anillos aleatorios, pero que este sesgo está explicado en buena medida por las restricciones existentes en el uso de las bases.

Para el anillo AB, la distancia con respecto al complemento es de 8, lo que lo deja dentro del 26 % con mayor palindromía en A_1, y el 20 % de mayor palindromía en A_0. Sin ser el más palíndrome, pertenece al menos a una minoría, de un 11 % si se tratase de secuencias aleatorias. El anillo AB nos interesa, por haber sido destacado en [Demongeot y Besson, 1996] como el más cercano a los loops del Gly-tRNA de *Œnothera*; sin embargo, hasta

[2] Al usar $\hat{d}_t$ entre nuestros anillos (en los cuales el uso de bases es muy similar), preferimos no considerar infinita la distancia cuando el uso de letras era diferente; en cambio, agregamos una condición de término alternativa, cuando ya nada más se puede extraer.

d_H^c	AZ	AZ_b	A_{uag}	A_{uga}	A_{uaa}	A_1	A_0
4	0.2	0.3	0.0	0.3	0.8	0.3	0.3
6	1.6	2.5	3.1	5.1	7.4	5.0	3.5
8	9.5	15.5	17.3	22.2	22.3	20.9	15.7
10	33.8	44.8	56.8	43.9	46.9	48.0	46.2
12	42.8	35.5	22.7	28.3	22.3	25.5	33.5
14	11.3	1.4	0.0	0.3	0.4	0.2	0.9
Promedio	11.1	10.3	10.2	10	9.9	9.7	9.9

Cuadro 6.6: Porcentaje de anillos con las distintas distancias d_H^c respecto de sus complementos.

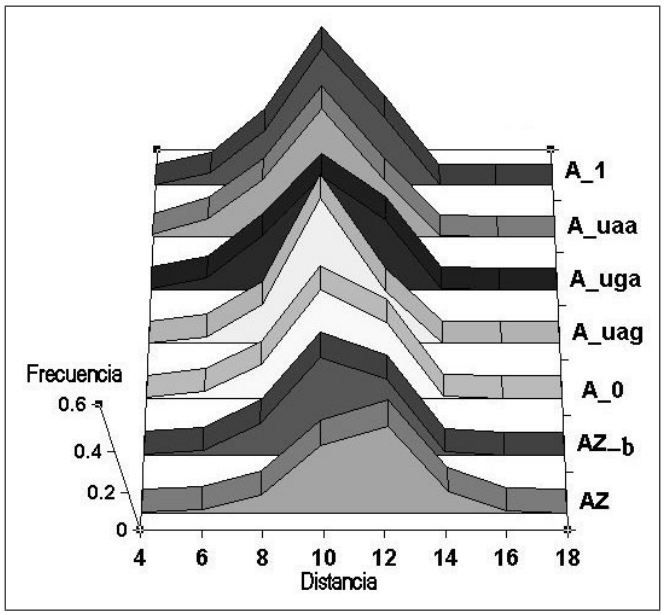


Figura 6.7: Histograma de distancia d_H^c respecto a los complementos de los anillos.

ahora no hay nada que lo singularice como candidato a "ganar el juego" primitivo, al menos no por sobre otros 334 anillos de A_1, ó 5723 de A_0.

Para ver un poco de la relación entre AB y los otros anillos, se calculó $d_H^c(\mathrm{AB}, x)$ para cada x en los distintos conjuntos usados arriba. Los resultados están graficados en la Figura 6.8; se puede ver que la distancia es menor en A_1 y A_0 que en los conjuntos aleatorios, incluso cuando se toma en cuenta la frecuencia de las bases. Sin embargo, era esperable que existiera una distancia menor, especialmente en A_1: las formas de codificar todos los aminoácidos con pocas bases no son muchas, por lo que es esperable que haya coincidencias de subcadenas entre AB y los otros anillos. De hecho, tal vez lo que debe extrañar es que la diferencia entre el promedio entre A_1 y las secuencias aleatorias no sea *mayor*.

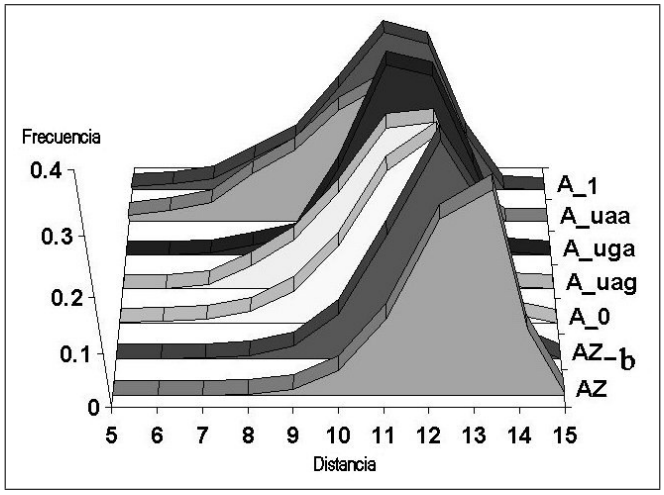


Figura 6.8: Histograma de distancia d_H^c respecto al anillo AB.

6.2.3. Formación de hairpins

Como dijimos antes, es altamente probable que existiera una ventaja selectiva, tal vez muy fuerte, a favor de cadenas que pudiesen tomar forma de *hairpin*. Esta forma del RNA es especialmente estable, y habría permitido a las secuencias sobrevivir sin ser destruidas, en momentos en que las condiciones ambientales se hicieran relativamente adversas (lo que pudo haber ocurrido con frecuencia incluso diaria, por el ciclo de días y noches, o con las mareas).

Un RNA en forma de hairpin lo que ha hecho es plegarse sobre sí mismo, formando un tallo (*stem*) mediante pareos de Watson-Crick entre sus propias bases. Las puntas que no alcancen a entrar en el tallo quedan libres, mientras que las bases situadas entre las dos subcadenas apareadas forman un bucle (*loop*). El loop debe tener al menos 3 bases.

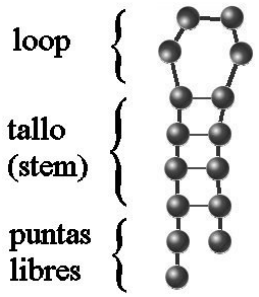


Figura 6.9: Partes de un hairpin.

Entre dos RNAs del mismo largo en forma de hairpin, será más estable aquel que tenga un tallo de mayor longitud, que tenga menos bases en las puntas libres, y que tenga un loop más pequeño (por sobre el mínimo de 3). También es favorable tener enlaces **g-c**

en el tallo, pues son más fuertes que **a-u**. Eventualmente se puede dar el pareo **g-u**, que es el más fuerte dentro de los pareos "no estándar"; en caso de darse, es preferible que quede *dentro* del tallo y no en uno de sus extremos, donde es más fácil que se abra. En particular, un hairpin con loop de 3 bases y una base libre es más estable que uno con loop de 4 bases y sin puntas libres. Todas estas afirmaciones, algunas de ellas en principio intuitivas, las hemos corroborado mediante cálculos de estructura secundaria usando el software RNAdraw de O. Matzura, versión 1.1 [Matzura y Wennborg, 1996].

Nos interesó, por lo tanto, ver cuáles eran los anillos capaces de formar mejores hairpins. El anillo AB de Demongeot, por ejemplo, forma un hairpin con tallo de largo 8, sin puntas libres, con un loop de 6 bases y un pareo no estándar en el extremo exterior del tallo.

Para cada anillo de A_1, A_0, y los dos conjuntos aleatorios, se calculó el mayor tallo posible, considerando todos los puntos de partida, todas las longitudes de tallo y todos los tamaños de loop mayores o iguales a 3, permitiendo un máximo de un pareo no estándar, del tipo **g-u**. La proporción de anillos que logró formar hairpins de distintos tamaños se indica en el Cuadro 6.7; la misma información se ha graficado en la Figura 6.10.

Tallo	AZ	AZ_b	A_0	A_1
2	2.31	0.01	0.00	0.00
3	24.18	11.45	8.16	1.72
4	43.36	46.38	47.63	32.42
5	21.73	29.30	30.79	34.38
6	6.37	9.52	10.17	21.41
7	1.60	2.60	2.48	6.95
8	0.35	0.63	0.65	2.19
9	0.09	0.12	0.12	0.94
Promedio	4.12	4.48	4.54	5.1

Cuadro 6.7: Porcentaje de anillos con tallos de los distintos tamaños.

Puede observarse cómo el tamaño de los tallos mejora al pasar de secuencias aleatorias a secuencias con la frecuencia de base de nuestros anillos, y cómo A_0 sobrepasa por poco la distribución de tamaños de estas últimas; A_1, por otro lado, tiene una clara tendencia hacia hairpins con mayores tallos, incluyendo una fracción 9 veces mayor de hairpins con tallo de largo 9 (lo mejor posible con secuencias de largo 22). El anillo AB, con su loop de tamaño 8, se ubica en un honroso 3 % de A_1, aunque con la desventaja de tener un loop de tamaño 6.

Como comentario, es necesario hacer notar que existe una correlación negativa entre el tamaño de los mejores tallos y la métrica d_H^c de los anillos con sus complementos; esta es de -0.45 para las secuencias aleatorias, -0.35 para las secuencias aleatorias con frecuencias de bases prefijadas, -0.27 en A_0 y -0.40 en A_1. Es razonable, pues la presencia de un hairpin *implica* una alta palindromía: si una subcadena α se parea con otra, β, dentro de la

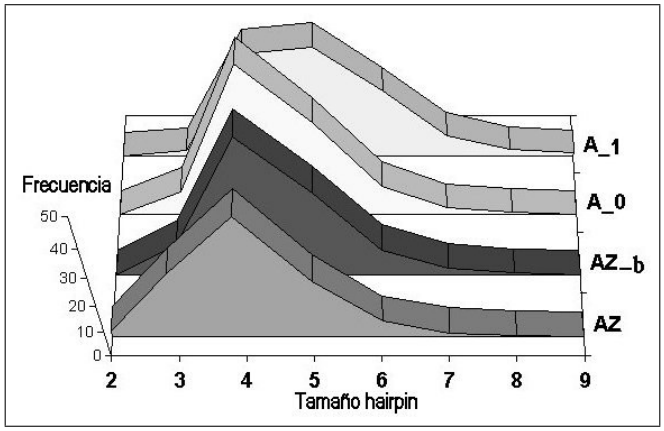


Figura 6.10: Histograma de tamaño de tallos en hairpins.

misma secuencia, significa que es idéntica al complemento de β; igualmente, β es idéntica al complemento de α, y por lo tanto, un tallo de largo k (suponiéndolo sin pareos no estándar) implica $2k$ bases idénticas entre el anillo y su complemento. Otra observación, trivial, es que para el anillo complementario existirá un tallo equivalente al del anillo original (con la sola salvedad de que en caso de haber un pareo **g-u**, en el complementario el tallo tendrá uno **c-a**, menos fuerte, aunque igualmente posible).

6.2.4. A*: Selección de 24 anillos

Basándonos en las razones para atribuir mayor probabilidad de supervivencia en el juego primitivo a los anillos con mejores hairpins, nos concentraremos desde aquí en los anillos de A_0 que presentan tallos de largo 9 y loops de tamaño 3. Recordemos que los extraemos de A_0, y no de A_1, debido a que este último conjunto es un subconjunto de A_0, separado en razón de la forma de sus secuencias, que es **aug,**... **,stop**. Este criterio, empleado en [Demongeot y Besson, 1996], es arbitrario, pues impone a las secuencias la forma que hoy conocemos como característica de un gen, en circunstancia de que en el escenario planteado, no hay nada que pueda favorecer esta forma *a priori*. A_0, en cambio, aparece naturalmente de la condición de codificación de las 21 clases del código genético en una longitud mínima.

El conjunto formado por los anillos de A_0 con tallo de largo 9 y loop de tamaño 3 lo denotaremos A^*, y consta de 24 elementos, listados en el Cuadro 6.9; en todos ellos el tallo está formado por 5 enlaces débiles, 3 fuertes y uno no estándar (**g-u**). Hay varias características notables en este conjunto: la metionina, que ya en A_0 resultaba ser el aminoácido más repetido por los anillos, sube aquí de 52 % a 79 % de repetición. Además, el 37.5 % de las secuencias tiene la forma **aug,**... **,stop**. En otras palabras, en lugar de imponer la repetición de M y la forma **aug,**... **,stop**, como se había hecho en [Demongeot y Besson, 1996], ahora vemos aparecer estas características por razones completamen-

te independientes: *emerge* la forma de los genes contemporáneos, y este es uno de los resultados más notables con los que nos encontramos a lo largo de esta investigación.

Es deber mencionar que también se observa un aumento (con respecto a A_0 ó A_1) del uso del stop **uga**. Si bien en los mamíferos este es el stop preferido, no lo es en las plantas, los invertebrados, ni en las bacterias, según nuestros cálculos realizados sobre la base de datos CUTG [Nakamura et al., 2000]. Por lo tanto, sólo cabe suponer que, en caso de que haya sido el codón predominante en el sistema primitivo, luego dejó de serlo por alguna razón evolutiva (por ejemplo, por ser el más proclive a errores en la transcripción [Lewin, 2002]).

Conjunto	Frecuencia	Promedio clases	Frecuencia	Frecuencia		
	aug,... ,stop	en complemento	repet. M	**uga**	**uaa**	**uag**
A_1	1.00	16.74	1.00	0.53	0.20	0.28
A_0	0.07	16.77	0.52	0.51	0.26	0.23
A^*	0.38	16.58	0.79	0.62	0.25	0.13
AZ	0.02	12.85	-	-	-	-
AZ_b	0.05	13.94	-	-	-	-

Cuadro 6.8: Algunas cantidades en distintos conjuntos de anillos. La forma **aug,... ,stop**, impuesta para A_1, emerge para A^*; la cantidad de clases codificadas por los anillos complementarios baja ligeramente en A^*; M resulta ser el aminoácido "sobrante", por repetición, y esto, junto a su posición respecto a **stop**, lo deja como candidato a **start**.

6.3. Frecuencias de subcadenas y selección de anillos óptimos

En esta sección nos reduciremos al conjunto A^* recién determinado, y supondremos que luego de ser seleccionados por los factores que se han explicado (codificación de las 21 clases del código, y mejor hairpin), estos anillos se encontraban en igualdad de condiciones para enfrentar el juego ya mencionado al comienzo de 6.2: un entorno en que en virtud de los ocasionales accidentes, las cadenas tienden a romperse, y eventualmente volver a formarse a partir de los trozos que las rodean.

¿Qué anillos llevarán ventaja en ese escenario? La intuición nos dice que serán aquellos que compartan más subcadenas con los demás, o, más en general, que se parezcan más a los demás, en los sentidos descritos por las distancias propuestas en 6.2.1 (y que reflejan el tipo de accidente que puede ocurrir a los anillos: errores en la replicación -o pareos con errores-, ruptura y rearme). Al ser más parecidos a los otros anillos, habrán más cadenas libres que les permitan rearmarse. De hecho, es posible pensar en una retroalimentación positiva, que a partir de una "comunidad" de anillos de A^*, en proporciones parejas, favorezca a los que se parecen a más de sus congéneres, aumentando en el tiempo su

	Secuencia	d_H^c c/r comp.	Clases en comp.	¿Sobra M?	¿Forma **aug,**... **,stop**?
1	aauucaugccagacugguauga	4	17	x	
2	caugccagaaauucugguauga	4	18	x	
3	auggugccacuauucaagauga	6	14	x	x
4	augcuauucaccaagaugguga	6	14	x	x
5	auggugcuaccauucaagauga	6	15	x	x
6	auggccuauucacaagauguga	6	15	x	x
7	augccacugguauucaagauga	6	16	x	x
8	auguggccuacauucaagauga	6	16	x	x
9	augccaagaugguauucacuga	6	16	x	x
10	gcaauguuuauggagaccauaa	6	16	x	
11	uauguuuggagaccaagcauaa	6	16		
12	uucaugccagaaacugguauga	6	17	x	
13	augguacugccauucaagauga	6	17	x	x
14	gcagaauguuuauggaccauaa	6	17	x	
15	aauauguuuggaccaagcauag	6	17		
16	uauguuuggaagccagacauaa	6	17		
17	uacauuuggaagccagauguaa	6	17		
18	acaauguuuauggaagccauag	6	17	x	
19	auggaagccauuuacaauguag	6	17	x	x
20	uacagauggaagccauuuguaa	6	17		
21	aacaugccagauucugguauga	6	18	x	
22	ugccagaaacauucugguauga	6	18	x	
23	uaugguucugcaagaaccauga	6	18	x	
24	uaccauucugcaagaaugguga	6	18		

Cuadro 6.9: Los 24 anillos de A*.

presencia, lo que a su vez mejora aún más su ventaja en el juego (pues habrá más cadenas de su propio tipo en el entorno).

No es un juego que se preste fácilmente para las consideraciones analíticas, y por esto llevamos a cabo una simulación, intentando entender mejor la dinámica de las cadenas rompiéndose y reuniéndose. Si bien algunos resultados preliminares parecen confirmar el panorama descrito en el párrafo anterior (apreciándose ventaja para los anillos más parecidos, en promedio, al resto del conjunto), nuestro esquema está aún en pañales, y me abstendré de presentarlo aquí.

Por el momento nos vemos obligados a suponer que los ganadores, dentro de A*, serán los anillos más cercanos, en promedio, al resto de A*; es una hipótesis razonable, a la espera de que una simulación acabada o algún otro tipo de análisis nos permita reforzar el punto.

¿Y qué noción de semejanza tomar? Para eso disponemos de las tres funciones de distancia presentadas en 6.2.1, tabuladas en los Cuadros 6.10 y 6.11. Nos interesa, según lo dicho

d_s \ d_H^c	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	•	8	13	13	13	12	12	11	8	13	12	9	12	13	13	11	12	12	12	12	10	11	13	10
2	**13**	•	13	12	13	13	13	12	12	13	12	9	12	13	11	11	12	12	13	13	8	8	12	12
3	**17**	**17**	•	10	2	11	6	5	14	10	12	11	4	12	10	12	13	11	11	12	11	11	11	7
4	**17**	**18**	**15**	•	10	11	9	11	10	12	13	11	9	11	13	13	12	10	9	12	12	13	12	9
5	**17**	**18**	**4**	**15**	•	12	6	5	13	12	12	11	2	12	9	13	13	12	11	12	11	10	11	5
6	**17**	**18**	**15**	**14**	**15**	•	13	8	10	12	11	13	12	9	10	12	11	12	12	12	14	13	12	13
7	**14**	**15**	**7**	**15**	**8**	**15**	•	7	10	12	12	10	4	11	11	12	10	11	9	7	9	11	10	11
8	**17**	**18**	**8**	**15**	**7**	**15**	**8**	•	11	11	12	12	6	10	11	12	12	11	10	12	11	12	12	10
9	**16**	**16**	**15**	**12**	**15**	**15**	**13**	**15**	•	12	13	13	11	13	13	12	11	11	11	12	13	12	12	11
10	**18**	**18**	**18**	**18**	**17**	**18**	**18**	**17**	**18**	•	8	12	12	8	11	9	9	4	7	9	13	11	12	12
11	**18**	**18**	**18**	**16**	**18**	**18**	**18**	**18**	**17**	**13**	•	12	12	8	7	4	8	9	10	10	11	11	9	13
12	**10**	**11**	**17**	**17**	**17**	**17**	**14**	**17**	**16**	**18**	**18**	•	12	12	12	11	12	11	12	11	4	8	12	13
13	**17**	**17**	**6**	**15**	**4**	**15**	**8**	**7**	**15**	**18**	**18**	**17**	•	11	10	12	12	12	10	11	10	12	10	7
14	**18**	**17**	**17**	**17**	**17**	**17**	**17**	**16**	**17**	**10**	**16**	**17**	**17**	•	8	10	12	8	10	13	13	12	11	13
15	**18**	**18**	**18**	**16**	**18**	**18**	**18**	**18**	**17**	**16**	**10**	**18**	**18**	**15**	•	8	10	9	12	11	11	12	11	12
16	**15**	**16**	**18**	**19**	**19**	**18**	**18**	**18**	**18**	**16**	**7**	**16**	**18**	**16**	**10**	•	4	6	8	8	10	8	11	14
17	**16**	**16**	**17**	**17**	**17**	**16**	**17**	**16**	**17**	**18**	**16**	**16**	**17**	**18**	**16**	**9**	•	8	8	4	12	11	11	12
18	**18**	**18**	**18**	**18**	**18**	**18**	**18**	**17**	**18**	**9**	**16**	**18**	**17**	**10**	**16**	**13**	**13**	•	4	8	13	10	11	12
19	**18**	**18**	**17**	**16**	**17**	**17**	**17**	**17**	**16**	**16**	**18**	**18**	**16**	**17**	**18**	**13**	**13**	**11**	•	4	13	10	13	12
20	**18**	**18**	**17**	**16**	**17**	**17**	**17**	**17**	**16**	**17**	**17**	**18**	**16**	**17**	**18**	**13**	**13**	**11**	**7**	•	11	12	12	11
21	**11**	**10**	**17**	**18**	**17**	**18**	**15**	**17**	**16**	**18**	**18**	**13**	**17**	**18**	**18**	**16**	**15**	**18**	**18**	**17**	•	7	11	12
22	**13**	**10**	**17**	**18**	**17**	**18**	**15**	**16**	**16**	**18**	**18**	**11**	**17**	**17**	**18**	**16**	**16**	**18**	**18**	**18**	**10**	•	10	10
23	**18**	**17**	**17**	**17**	**17**	**17**	**17**	**17**	**17**	**17**	**17**	**17**	**17**	**17**	**17**	**17**	**19**	**17**	**18**	**18**	**17**	**17**	•	6
24	**18**	**16**	**14**	**15**	**14**	**16**	**17**	**17**	**17**	**17**	**18**	**18**	**15**	**16**	**17**	**19**	**18**	**18**	**17**	**17**	**16**	**15**	**11**	•

Cuadro 6.10: Distancia entre los anillos de A*: arriba a la derecha, d_H^c; abajo a la izquierda, en negritas, d_s (los ceros de la diagonal -de ambas- se omiten).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	0	2	7	9	7	9	7	8	8	10	9	2	5	10	10	7	9	11	11	10	2	5	10	9
2		0	8	9	10	9	8	9	8	11	10	2	7	10	9	7	8	10	11	10	2	3	8	7
3			0	5	3	5	3	5	6	12	11	8	2	11	10	10	9	11	9	10	8	8	8	8
4				0	7	4	5	7	3	9	10	9	8	8	9	12	11	11	10	9	9	9	9	7
5					0	7	5	4	8	10	10	9	3	10	10	11	10	11	9	10	8	10	8	6
6						0	7	5	5	12	11	10	7	10	8	9	8	12	9	9	9	9	8	8
7							0	4	3	10	10	6	3	9	10	11	8	10	10	10	7	8	9	9
8								0	8	10	12	9	5	10	10	10	8	10	8	10	9	8	10	9
9									0	10	10	7	6	8	9	11	9	10	9	8	7	8	9	8
10										0	5	10	11	3	6	8	9	3	7	9	10	11	9	10
11											0	9	10	6	3	3	6	7	10	8	9	10	9	11
12												0	6	9	9	7	8	11	10	9	3	2	8	8
13													0	10	10	10	10	10	7	8	7	9	7	6
14														0	7	7	9	5	7	7	10	11	8	10
15															0	5	8	6	8	10	10	10	8	9
16																0	3	5	8	5	7	8	10	11
17																	0	7	5	3	7	8	12	11
18																		0	3	7	10	9	8	10
19																			0	3	10	10	10	10
20																				0	10	9	11	9
21																					0	3	9	10
22																						0	8	9
23																							0	3
24																								0

Cuadro 6.11: *Distancia entre los anillos de A*: $\hat{d}_t$.*

arriba, ver la distancia *promedio* de cada anillo con respecto al resto de A*; estos promedios se presentan en el Cuadro 6.12.

El Cuadro 6.13 muestra las posiciones de los distintos anillos en el ranking de cada distancia; claramente, hay unos pocos anillos (especialmente el número 13) que se destacan en los tres casos. Esto era esperable, pues existe una alta correlación entre las tres distancias: si hay una subcadena grande compartida por dos anillos, la distancia de Hamming ya parte de una base alta, y también serán menos los cortes necesarios para convertir un anillo en el otro. Esta relación se ve especialmente exacerbada en A*, por el alto grado de

Anillo	d_H^c	d_s	$\hat{d}_t$
1	11.04	15.50	7.38
2	11.13	15.46	7.42
3	9.67	14.33	7.38
4	10.71	15.58	7.83
5	9.58	14.29	7.75
6	11.17	15.92	7.92
7	9.42	14.13	7.13
8	9.75	14.50	7.79
9	11.17	15.33	7.38
10	10.17	15.88	8.42
11	10.04	15.79	8.21
12	10.54	15.38	7.21
13	9.38	14.25	7.08
14	10.54	15.58	8.17
15	10.21	16.00	8.13
16	9.63	14.92	7.79
17	9.96	15.25	7.96
18	9.46	15.25	8.08
19	9.63	15.46	8.08
20	9.96	15.42	8.21
21	10.42	15.33	7.29
22	10.21	15.29	7.63
23	10.63	16.25	8.42
24	10.29	15.67	8.29

Cuadro 6.12: *Promedio de las distancias de los anillos, en A^*.*

Posición	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
d_H^c	13	7	18	5	16	19	3	8	17	20	11	10	15	22	24	21	12	14	23	4	1	2	6	9
d_s	7	13	5	3	8	16	17	18	22	9	21	12	20	2	19	1	4	14	24	11	10	6	15	23
$\hat{d}_t$	13	7	12	21	1	3	9	2	22	5	8	16	4	6	17	18	19	15	14	11	20	24	10	23
Ranking promedio	13	7	3	5	16	8	18	17	21	12	22	19	9	1	2	20	4	11	15	14	10	24	6	23

Cuadro 6.13: *Ranking de los anillos de acuerdo a las diferentes distancias promedio (con la menor distancia arriba), además de sus posiciones en un ranking promediado (se les ubicó según el promedio de su posición en las tres distancias).*

palindromía de las secuencias.

De hecho, lo que ocurre es que, como veremos a continuación, los "mejores" anillos son muy parecidos entre sí (para los anillos **13** y **5**, por ejemplo, d_H^c es sólo 2), y ocupan el "baricentro" de A^*. Los anillos **13**, **7**, **3**, **5** y **8** pertenecen a este grupo. Sólo **16** se entromete entre ellos, gracias a una posición central en otro conjunto de anillos relativamente similares (entre sí). Lo que estamos viendo es la presencia de una estructura de *clusters*, así que a ella nos abocaremos a continuación.

La técnica usada más comúnmente para generar un gráfico de clusters a partir de una tabla de distancias es el *clustering jerárquico*, que parte agrupando elementos sueltos, y luego

trata los clusters resultantes como nuevos elementos, y aplica las reglas sucesivamente hasta reunir todos los elementos en un árbol. Existen diversas versiones, que se diferencian básicamente por la forma en que definen la distancia entre un cluster y otro. Las tres opciones más comunes son *single linkage*, *complete linkage* y *average linkage*. En *single linkage*, se define la distancia entre dos clusters como el *mínimo* de las distancias entre los elementos de uno y otro; el resultado suele ser la aparición de clusters conectados por un solo elemento, y en general, clusters con forma de cadenas. En *complete linkage* se hace todo lo contrario: se toma el *máximo* de las distancias entre los elementos de los dos clusters. Este método suele ser apropiado cuando existen clusters relativamente naturales en los datos. Finalmente, *average linkage* define la distancia entre dos clusters como el *promedio* de las distancias entre todos los pares de elementos; se le suele considerar robusto tanto ante situaciones de "cadenas" como de clusters naturales.

Escribimos un pequeño programa para aplicar estos tres métodos sobre las tablas de distancias entre los anillos[3]; el resultado son los agrupamientos esquematizados mediante paréntesis en el Cuadro 6.14.

Distancia	Método	Estructura de clusters
d_H^c	Average	(((1,2),((12,21),22)),(((((((3,5),13),7),8),(23,24)),(4,9)),((6,(((11,16),15),14)),(((10,18),19),(17,20)))))
	Linkage	(((((((1,2),9),((12,21),22)),(((((((3,5),13),7),8),24),23),((((((10,18),19),20),17),16),11)),15)),6),14),4)
	Complete	((((1,2),((12,21),22)),(((((3,5),13),7),8),(23,24))),(((4,9),(((10,18),19),(17,20))),((6,14),((11,16),15))))
d_s	Average	((((1,12),((2,21),22)),(((((((3,5),13),8),7),(4,9)),6),(23,24))),((((10,18),14),((11,16),15)),(17,(19,20))))
	Linkage	((((((1,12),((2,21),22)),(((((3,5),13),7),8),(4,9))),6),(23,24)),((((10,18),14),(19,20)),(((11,16),17),15)))
	Complete	(((((1,12),((2,21),22)),(((((((3,5),13),7),8),(4,9)),6),(17,(19,20)))),(((10,18),14),((11,16),15))),(23,24))
$\hat{d}_t$	Average	((((((1,2),12),21),22),((((3,13),5),7),8),((4,9),6))),(23,24)),((((10,14),(18,19)),(11,15)),((16,17),20)))
	Linkage	((((((1,2),12),21),22),(((((((3,13),5),7),9),4),6),8)),(23,24)),(((((((10,14),18),19),20),17),16),11),15))
	Complete	((((((1,2),12),21),22),((((3,13),5),(7,8)),((4,9),6))),(23,24)),(((10,14),(11,15)),(((16,17),20),(18,19))))

Cuadro 6.14: Clusters formados por los anillos, bajo distintos métodos de clustering jerárquico, para las diferentes distancias.

[3]La única salvedad es que usamos una versión simplificada de *average linkage*, sin memoria de los elementos que se han unido en un cluster: cada vez que se unen dos clusters, definimos la distancia del cluster resultante, respecto a otros clusters, como el promedio de las que tenían los dos clusters que se unieron.

Puede observarse que la estructura de clusters es bastante robusta, y aparece de manera similar para las distintas distancias y técnicas de clustering (nótese que algunas similitudes pueden estar escondidas por simples cambios de orden en la escritura de las uniones). Una forma de graficar los clusters es mediante *dendrogramas*, que además representan la historia del proceso de clustering (que no se ve en el esquema final)[4]; en la Figura 6.11 se muestra un dendrograma para el *average linkage* de d_s. En el se puede apreciar, por ejemplo, que el conjunto de "ganadores" en el ranking (**3, 5, 7, 8, 13**) constituyen el cluster principal dentro del cluster principal dentro del cluster principal, etc.; mientras que **16**, que logró quedar 5to en el ranking, ocupa una posición similar (en este y en otros esquemas del Cuadro), pero dentro de un gran cluster secundario. También es posible notar que, salvo **19**, todos los anillos de la forma **aug,**... **,stop** están juntos en un mismo cluster; además, la gran rama izquierda está formada completamente por anillos que repiten el aminoácido M (salvo **24**, que está en un extremo). En particular, el cluster ganador del "juego" está formado por anillos con ambas características.

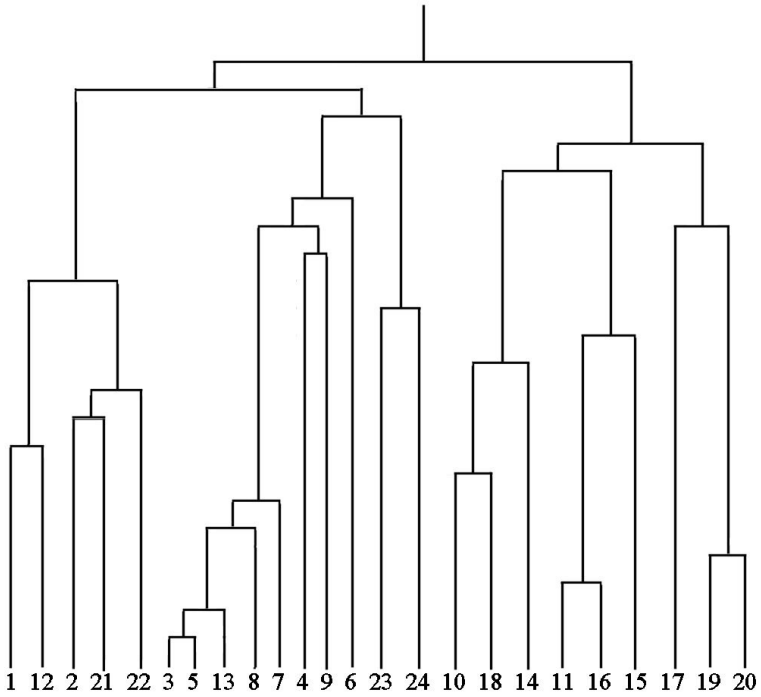


Figura 6.11: Dendrograma para los clusters formados en (A^*, d_s) mediante *average clustering*.

[4]Además es posible alargar las líneas verticales de acuerdo a las distancias entre los clusters; esto, que es muy útil por ejemplo al elaborar árboles filogenéticos, no se ha hecho en nuestro ejemplo.

Capítulo 7

Relación con secuencias reales

En las secciones anteriores todo lo que se ha hecho, tanto en simulaciones como en determinación y selección de conjuntos de anillos, ha sido sobre la base de modelos y criterios razonables sobre cómo pudieron darse las condiciones en la "sopa" o "pizza" primitiva: no ha habido ningún dato biológico, nada de secuencias actualmente existentes en la biología. En este capítulo exploraremos esa relación. Primero que nada, intentando (con éxito sólo relativo) ver si las secuencias de los anillos están estadísticamente relacionadas, a través de la frecuencia con que aparecen los dinucleótidos y codones, con los genomas actuales. En segundo lugar, y con resultados bastante más interesantes, veremos la relación de los anillos con las partes libres de los tRNAs contemporáneos, los "fósiles moleculares" más confiables que se conocen, y directamente ligados a la historia que surge del juego primitivo.

7.1. Los anillos como secuencias primitivas

Existen diversas cantidades que han sido calculadas en base a las frecuencias de bases, dinucleótidos o codones en los genes contemporáneos. Es interesante evaluar tales estadísticas en los anillos de A_1 y A_0, pues si se supone que estos anillos constituyeron el primer material genético (siendo integrados y modificados en los primeros genes), podría ser que su ubicación respecto a los reinos de los seres vivos confirmase esa posición primitiva. Sin embargo, es importante destacar que esto no es un test que pueda reafirmar o negar la teoría en su conjunto; muchos autores han considerado muy difícil, o imposible, que las estadísticas de los genes contemporáneos, después de más de 3 mil millones de años de deriva, selección y acción de diversos operadores genéticos, conserven alguna marca de las secuencias originales, cualesquiera que éstas hayan sido. Hacemos estos cálculos sólo porque los resultados *podrían* ser de interés.

Lo primero que llama la atención respecto a las frecuencias de letras y palabras cortas en A_0 (y en sus subconjuntos A_1 y A^*) es el hecho de que sólo existen dos patrones

de uso de bases y de dinucleótidos (Cuadro 6.2), dependiendo de cuál sea el codón de término utilizado. Llama más la atención al ver que esa rigidez va acompañada del uso de 50 codones distintos. Incluso al restringirnos a A^*, un conjunto de sólo 24 secuencias, muchas de las cuales comparten largas subcadenas, siguen usándose 45 codones distintos, y las frecuencias resultan representativas de A_1, un conjunto mil veces mayor.

Otro aspecto que llama la atención es que sólo un dinucleótido está ausente de los anillos, y resulta ser **cg**, que es también el dinucleótido más escaso en los genes de *Archaea* y *Eukaryota* [Karlin y Mrázek, 1997]. ¿Podría considerarse esa escasez como un vestigio de la ausencia de **cg** en el material genético más antiguo? Es eso lo que se sugiere en [Demongeot y Besson, 1996]; sin embargo, también se han propuesto otras razones para explicar esa escasez en los genomas actuales, y es una discusión que escapa al alcance de este texto.

7.1.1. Índice de homogeneidad local

Una cantidad que fue propuesta por P. Miramontes y sus colaboradores es el IDH, *Index of DNA Homogeneity* [Miramontes et al., 1995], que pretende medir el nivel de homogeneidad en secuencias binarias. Se trata de tres índices (sólo dos –cualesquiera– de ellos realmente independientes), cada uno generado por la proyección de la secuencia en uno de los tres alfabetos binarios: R/Y (con R={**a**, **g**}, Y={**c**, **u**}), W/S (con W={**a**, **u**}, S={**c**, **g**}), y M/K (con M={**a**, **c**}, K={**g**, **u**}). Si llamamos a las letras del alfabeto binario 0 y 1, entonces el IDH es

$$d \;=\; \frac{\rho_{00}\rho_{11} - \rho_{01}\rho_{10}}{\rho_0\rho_1}$$

donde cada ρ_α indica la frecuencia de ocurrencia de α en la secuencia. De esta manera, secuencias con mucha alternancia de 0 y 1 tendrán un IDH muy negativo, mientras que secuencias con largos tramos de 0 o de 1 tendrán IDH muy positivo; para secuencias aleatorias será cercano a 0. En [Miramontes et al., 1995] se mapean diversas especies al plano d_{ry}, d_{ws}; se observa que los genes de una misma especie tienden a caer en un mismo sector del plano, y más aún, que los distintos reinos o phyla tienden a ocupar zonas distintas de éste. En [Moreira, 1999] y [Moreira, 2003] se repitieron estos cálculos para una serie de especies, y si bien los resultados no fueron tan buenos como los expuestos en [Miramontes et al., 1995] (y en comunicación personal, Miramontes reconoció que ellos también habían notado que el mapeo no era tan bueno), sí pudo constatarse que el IDH constituía un buen indicador (junto con otros) de las diferencias entre los "estilos genómicos" de las distintas especies y reinos.

Calculamos d_{ry} y d_{ws} para A_{uaa} y A_{uga} (a partir de los Cuadros 6.1 y 6.2). Como A_1 y A_0 están formados por anillos que tiene la distribución de dinucleótidos de uno u otro de estos grupos (según el stop usado), su IDH promedio se obtiene por la combinación lineal adecuada de los anteriores. Los resultados están en la Figura 7.1 (adaptada de [Moreira, 2003]).

Lo que se observa es que los anillos quedan ubicados en la zona "bacteriana", con un valor

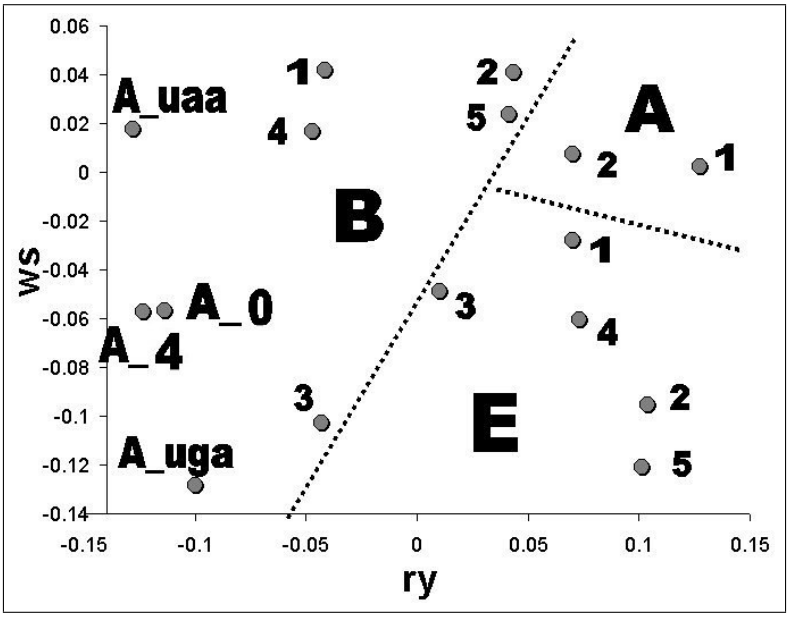


Figura 7.1: IDH (proyecciones R/Y y W/S) para A_{uaa}, A_{uga}, A_1, A_0 y diversas especies de *Eubacteria* (B1–B5), *Archaea* (A1, A2), y *Eukarya* (E1–E5).

d_{ry} menor a todas las especies graficadas. Es difícil interpretar la distancia de los puntos, ya que, como se mencionó en 3.2.3, no existe una forma comúnmente aceptada para la forma del árbol de la vida, a nivel primitivo, que relacione a los tres reinos (y de hecho, dicho árbol puede simplemente no existir). No es posible, por lo tanto, concluir nada de aquí; pero si adscribiéramos a la tesis que señala que las eubacterias (**B** en este gráfico) son los seres más antiguos, y además creyésemos que sus secuencias actuales reflejan sus secuencias antiguas, entonces podríamos ver en la ubicación de los anillos una muestra de su cercanía con éstas.

7.1.2. GC y redundancia

Otro cálculo, realizado por Demongeot en [Demongeot y Besson, 1996], hace uso de la "curva de Gatlin" [Gatlin, 1968], que ubica las secuencias en un plano poniendo en un eje el contenido de *gc* (es decir, la suma de las frecuencias de **g** y **c**), y en el otro la "redundancia" I, definida como

$$I \;=\; 1 + p\,\{p_1 \log_2 p_1 + (1-p_1)\log_2(1-p_1)\} + (1-p)\,\{p_2 \log_2 p_2 + (1-p_2)\log_2(1-p_2)\}$$

Si proyectamos **c**, **g** $\to 0$, **a**, **u** $\to 1$, entonces p es la frecuencia de 0 (es decir, *gc*), p_1 es la probabilidad de que un 0 siga a un 0 (es decir, $p_1 = \rho_{00}/\rho_0$), y p_2 es la probabilidad de que un 0 siga a un 1 (es decir, $p_2 = \rho_{10}/\rho_1$). En la Sección 4.2 se mostró el cuadro obtenido por Demongeot (Figura 4.7). Ahí Demongeot ubicaba su anillo AB; es decir, A_{uga} (pues todos los anillos de ese conjunto comparten el mismo uso de dinucleótidos).

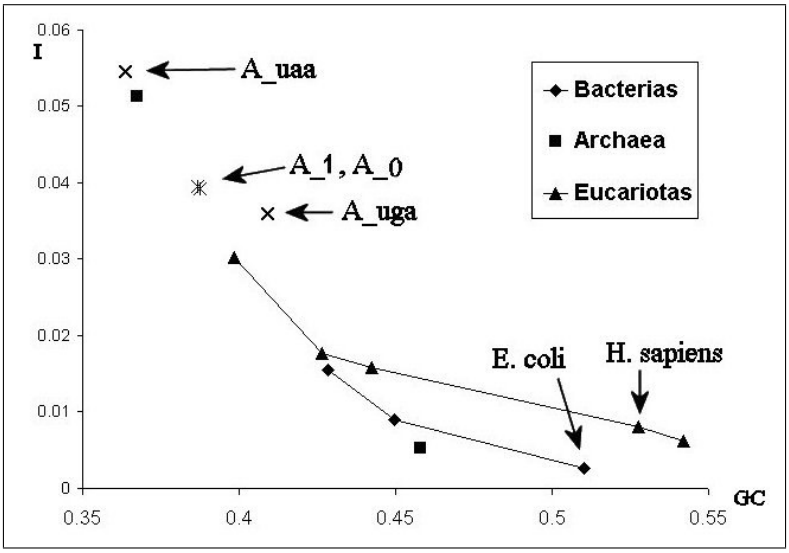


Figura 7.2: Ubicación de los anillos y algunos genomas en el plano *gc* vs. I.

En la Figura 7.2 hemos ubicado A_{uga}, A_{uaa}, así como A_1 y A_0 (que son combinaciones lineales de los primeros). Además, hemos usado datos "frescos" para ubicar distintas especies en el plano. Los datos, tomados de [Moreira, 2003] y sus materiales complementarios, provienen de la versión 131 de Genbank [Benson et al., 2002], y son por lo tanto bastante más masivos que los pocos usados por Demongeot en 1996 (y que provenían de 1968 y 1990). Corresponden a dos especies de arqueobacterias, 3 de bacterias y 5 de eucariotas. Compárese esta figura con la Figura 4.7; al igual que aquella, ésta muestra a los genomas eucariotas un poco por sobre de los bacterianos. Sin embargo, ya no es tan clara la posición de AB (es decir, A_{uga}) con respecto a ellos, y no nos parece que se pueda concluir nada a partir de esta representación.

7.1.3. Uso de codones

Otra pregunta que cabe hacerse es respecto a la frecuencia de uso, en los genomas actuales, de los codones usados por los anillos de A_0. En [Demongeot y Besson, 1996] se intentó estudiar esta posible relación para A_1, pero los datos no eran muchos. Hoy en día el uso de codones se ha tabulado para muchas especies [Nakamura et al., 2000], basándose en una gran cantidad de secuencias; además, podemos hacer la comparación con todo A_0, y no sólo con un anillo particular.

En el Cuadro 7.1 se ha tabulado el uso relativo de codones en A_0 y en 12 genomas representativos, para 18 aminoácidos, usando información disponible en los materiales complementarios de [Moreira, 2003] (se omiten los codones de término, para los que faltaba la información, y también M y W, que sólo se codifican con un aminoácido). El uso relativo (RSCU, por *relative synonymous codon usage*) se calcula como la frecuencia con que un

codón es utilizado para codificar un cierto aminoácido, multiplicada por la cantidad de codones que lo hacen (de modo de poder comparar el grado de uso de codones en clases de equivalencia de distinto tamaño). Al calcularse las correlaciones entre RSCU para A_0 y para cada una de las especies, se encontró correlación positiva en 8 de los 12 casos (y cercana a 0 en otros dos; sólo en 2 fue claramente negativa).

Además, se calculó para cada codón la desviación estándar de sus valores entre las 12 especies; los valores resultantes tuvieron también una correlación positiva, aunque baja (0.53), con la diferencia entre el RSCU promedio de las especies con el RSCU de A_0, lo que indica que en los casos en que el uso de los codones es similar en el conjunto de las especies, éste también es similar con respecto a A_0. Para A^*, el selecto conjunto de 24 anillos en 6.3, el uso de codones es muy similar al de A_0, y las correlaciones, aunque ligeramente menores, mantienen las mismas características. Todo esto podría interpretarse como una cercanía no trivial entre la ocurrencia de los codones en los anillos "primitivos" y su ocurrencia en los genomas actuales.

Nota: Las especies usadas en el Cuadro 7.1 y también en las Figuras 7.1 y 7.2 son las siguientes: Methanosarcina acetivorans C2A ($A1$), Sulfolobus solfataricus ($A2$), Escherichia coli ($B1$), Bacillus subtilis ($B2$), Streptomyces coelicolor A3(2) ($B3$), Mesorhizobium loti ($B4$), Nostoc sp. PCC 7120 ($B5$), Saccharomyces cerevisiae ($E1$), Arabidopsis thaliana ($E2$), Drosophila melanogaster ($E3$), Caenorhabditis elegans ($E4$) y Homo sapiens ($E5$). Fueron seleccionadas (para [Moreira, 2003]) de acuerdo a dos criterios: que estuvieran representadas por un alto número de secuencias, y que su conjunto estuviese bien distribuido dentro del árbol de la vida.

7.2. Datos de tRNAs contemporáneos

Hasta hace unos pocos años, la disponibilidad de información respecto a secuencias de tRNA aún era bastante baja; cuando se escribió el último artículo de J. Demongeot [Demongeot y Besson, 1996], era necesario buscar artículos específicos en que se describiera uno u otro tRNA. Hoy en día la situación ha cambiado, gracias a los proyectos de secuenciamiento masivo, y a la automatización de la gestión de la información, haciendo posible comparar grandes cantidades de secuencias. Entre otras cosas, esto ha permitido mejorar la información sobre las posiciones conservadas en el tRNA, y también ha hecho posible la estructuración de bases de datos específicamente dedicadas a él (e incluso a temas más específicos). Esto permite retomar la comparación de los anillos con las partes libres del tRNA, haciéndola ahora masiva y más representativa.

El primer paso para hacer esto fue encontrar un catálogo de secuencias de tRNA. El primer intento consistió en obtener los archivos completos de la versión más reciente de Genbank (la 131, en marzo del 2002), y usando una versión modificada del software TIP [Moreira, 2003; Moreira y Maass, 2003] extraer de ellas todas las secuencias de tRNA anotadas como tales. En primera instancia, esto parecía prometedor, pues aparecieron más de 20.000 entradas; además, esta utilización directa de Genbank hubiese permitido contar

A-á	Codón	A1	A2	B1	B2	B3	B4	B5	E1	E2	E3	E4	E5	A_0
A	gca	1.5	1.4	0.9	1.1	0.1	0.4	1.1	1.2	1.1	0.7	1.3	0.9	2
	gcc	1.1	0.5	1.1	0.8	2.3	1.9	0.9	0.9	0.6	1.8	0.8	1.6	1.7
	gcg	0.4	0.5	1.3	1.1	1.5	1.4	0.6	0.4	0.5	0.8	0.5	0.4	0
	gcu	1	1.6	0.7	1	0.1	0.3	1.4	1.5	1.7	0.8	1.4	1	0.4
C	ugc	1.1	0.6	1.1	1.1	1.8	1.7	0.7	0.7	0.8	1.4	0.9	1.1	1.1
	ugu	0.9	1.4	0.9	0.9	0.2	0.3	1.3	1.3	1.2	0.6	1.1	0.9	0.9
D	gac	0.9	0.5	0.7	0.7	1.9	1.3	0.6	0.7	0.6	1	0.6	1.1	0.5
	gau	1.1	1.5	1.3	1.3	0.1	0.7	1.4	1.3	1.4	1	1.4	0.9	1.5
E	gaa	1.4	1.1	1.4	1.4	0.3	0.9	1.5	1.4	1	0.6	1.3	0.8	1.4
	gag	0.6	0.9	0.6	0.6	1.7	1.1	0.5	0.6	1	1.4	0.7	1.2	0.6
F	uuc	0.9	0.8	0.9	0.7	2	1.6	0.5	0.8	0.9	1.3	1	1.1	1
	uuu	1.1	1.2	1.1	1.3	0	0.4	1.5	1.2	1.1	0.7	1	0.9	1
G	gga	1.5	1.6	0.5	1.2	0.3	0.3	0.9	0.9	1.5	1.1	2.3	1	2
	ggc	0.8	0.4	1.5	1.4	2.6	2.8	0.8	0.8	0.5	1.7	0.5	1.4	1
	ggg	1	0.6	0.6	0.6	0.7	0.4	0.6	0.5	0.6	0.3	0.3	1	0
	ggu	0.7	1.4	1.4	0.7	0.4	0.5	1.7	1.9	1.3	0.8	0.8	0.6	1
H	cac	0.9	0.7	0.9	0.7	1.9	1	0.8	0.7	0.8	1.2	0.8	1.2	0.3
	cau	1.1	1.3	1.1	1.3	0.1	1	1.2	1.3	1.2	0.8	1.2	0.8	1.7
I	aua	0.8	1.6	0.3	0.4	0.1	0.2	0.5	0.8	0.7	0.6	0.5	0.5	1
	auc	1.1	0.3	1.2	1.1	2.9	2.4	0.8	0.8	1	1.4	0.9	1.5	0
	auu	1.1	1.1	1.5	1.5	0.1	0.4	1.7	1.4	1.2	1	1.6	1.1	2
K	aaa	1.3	1	1.5	1.4	0.1	0.4	1.5	1.2	1	0.6	1.2	0.8	0.8
	aag	0.7	1	0.5	0.6	1.9	1.6	0.5	0.8	1	1.4	0.8	1.2	1.2
L	cua	0.3	1.1	0.2	0.3	0	0.1	0.8	0.8	0.6	0.5	0.6	0.4	0.5
	cuc	1.2	0.4	0.6	0.7	2.2	1.6	0.6	0.3	1	0.9	1	1.2	0
	cug	1.6	0.3	2.9	1.5	3.6	2.9	0.7	0.7	0.6	2.6	0.8	2.4	2.2
	cuu	2	0.9	0.7	1.4	0.1	0.7	0.5	0.8	1.5	0.6	1.5	0.8	0.4
	uua	0.5	2.5	0.8	1.2	0	0	2.2	1.7	0.8	0.3	0.7	0.4	1.6
	uug	0.4	0.8	0.8	0.9	0.1	0.6	1.2	1.7	1.4	1.1	1.4	0.8	1.3
N	aac	1	0.7	1.1	0.9	1.9	1.3	0.7	0.8	0.9	1.1	0.7	1.1	0.5
	aau	1	1.3	0.9	1.1	0.1	0.7	1.3	1.2	1.1	0.9	1.3	0.9	1.5
P	cca	0.6	1.8	0.8	0.7	0.1	0.4	1.3	1.7	1.3	1	2.1	1.1	3.3
	ccc	1	0.5	0.5	0.3	1.7	1	1	0.6	0.4	1.3	0.4	1.3	0
	ccg	1	0.4	2.1	1.8	2.2	2.3	0.4	0.5	0.7	1.2	0.8	0.5	0
	ccu	1.4	1.3	0.7	1.1	0.1	0.3	1.2	1.2	1.5	0.5	0.7	1.1	0.7
Q	caa	0.5	1.5	0.7	1	0.1	0.4	1.4	1.4	1.1	0.6	1.3	0.5	1.2
	cag	1.5	0.5	1.3	1	1.9	1.6	0.6	0.6	0.9	1.4	0.7	1.5	0.8
R	aga	1.6	3.4	0.3	1.5	0.1	0.2	1	2.9	2.1	0.6	1.8	1.2	6
	agg	2	2.2	0.2	0.6	0.2	0.5	0.4	1.3	1.2	0.7	0.5	1.2	0
	cga	0.4	0.2	0.4	0.6	0.2	0.3	0.8	0.4	0.7	0.9	1.4	0.7	0
	cgc	0.7	0.1	2.2	1.3	2.9	3.2	1.5	0.3	0.4	2	0.6	1.2	0
	cgg	0.8	0.1	0.6	0.9	2.3	1.2	0.8	0.2	0.5	0.9	0.6	1.3	0
	cgu	0.5	0.2	2.3	1.1	0.4	0.6	1.5	0.9	1	0.9	1.3	0.5	0
S	agc	1.1	0.6	1.5	1.4	1.4	1.5	0.9	0.7	0.8	1.5	0.6	1.5	1.3
	agu	0.9	1.5	0.9	0.6	0.2	0.3	1.6	0.9	1	0.8	0.9	0.9	1.6
	uca	1.1	1.4	0.8	1.4	0.1	0.3	0.9	1.3	1.2	0.5	1.5	0.9	1.4
	ucc	1.3	0.7	0.9	0.8	2.5	1.3	0.7	1	0.7	1.4	0.8	1.3	1
	ucg	0.6	0.4	0.8	0.6	1.7	2.4	0.4	0.6	0.6	1.2	0.9	0.3	0
	ucu	1	1.4	1	1.2	0.1	0.2	1.5	1.6	1.7	0.5	1.2	1.1	0.6
T	aca	1.1	1.2	0.6	1.7	0.1	0.3	1.3	1.2	1.3	0.7	1.4	1.1	1.8
	acc	1.3	0.6	1.7	0.6	2.6	2	1	0.9	0.8	1.5	0.7	1.4	1.7
	acg	0.6	0.5	1	1.1	1.2	1.4	0.3	0.5	0.6	1.1	0.6	0.5	0
	acu	1	1.8	0.7	0.6	0.1	0.2	1.4	1.4	1.4	0.7	1.3	1	0.5
V	gua	1	1.6	0.7	0.8	0.1	0.1	1.2	0.8	0.6	0.4	0.7	0.5	1.4
	guc	1	0.4	0.8	1.1	2.2	2	0.7	0.8	0.7	0.9	0.9	1	0
	gug	0.8	0.6	1.4	1	1.6	1.4	0.9	0.8	1	1.9	0.9	1.9	1.6
	guu	1.2	1.5	1.1	1.1	0.1	0.4	1.2	1.6	1.6	0.7	1.6	0.7	1
Y	uac	0.9	0.7	0.8	0.7	1.9	1	0.7	0.9	0.9	1.3	0.9	1.1	0.6
	uau	1.1	1.3	1.2	1.3	0.1	1	1.3	1.1	1.1	0.7	1.1	0.9	1.4

Cuadro 7.1: Uso relativo de codones en A_0 y 12 especies (A1 y A2 son arqueobacterias, B1 a B5 son eubacterias, y E1 a E5 son eucariotas).

con información sobre los intrones existentes en muchos tRNA, probablemente relevantes para su historia antigua, y que no están presentes en otras fuentes. Por desgracia, pronto la información demostró ser de poca utilidad, por varias razones. En primer lugar, apare-

ció una gran cantidad de tRNAs de longitud anómala, desde una docena de bases, hasta varios miles (y un chequeo demostró que las anomalías estaban en los archivos originales, y no en el software utilizado). Además, los tRNAs no estaban bien anotados: muchas veces, en lugar de usar la entrada "tRNA" en la tabla de contenidos de una secuencia, se utilizaba "gene", y sólo se indicaba tRNA como "product"; en otros casos, sólo se indicaba que "dentro de esta secuencia hay un tRNA", sin precisar su ubicación. Tampoco había información alguna sobre la conformación bidimensional, o en general, sobre el alineamiento de la secuencia respecto de la forma estándar del tRNA. Finalmente, estaba el problema de trabajar con un listado generado específicamente, en lugar de usar alguno accesible para el resto de los investigadores y que hiciese repetibles los cálculos. Salvo este último inconveniente, los demás son parte de la organización de Genbank, que por lo visto no ha sido puesta a punto para un uso sistemático de la información sobre tRNAs. Finalmente, debí descartar el uso de los datos obtenidos, y favorecer en cambio el uso de una base de datos universalmente reconocida como referencia en el tema. Lamentablemente, eso implicó desechar también la idea de estudiar los intrones de los tRNA, hasta que se disponga de buena información sobre ellos.

La alternativa, entonces, fue utilizar la base de datos de tRNA presentada en [Sprinzl et al., 1998], que además de tener información de buena calidad, ha pasado a convertirse en el punto de referencia obligado en los estudios sobre tRNA. A pesar de la fecha de la publicación del artículo, la base de datos misma tiene información más actualizada (al menos hasta el año 2000). Está disponible tanto en archivos de texto como en planillas para Microsoft Excel, y para cada tRNA listado, se presenta el alineamiento, la estructura secundaria (indicando los pareos de las bases), y una referencia a la fuente de la secuencia. A partir de dicho listado, extractamos todas aquellas secuencias que no contenían bases distintas a **a**, **c**, **g**, **u**, y además eliminamos las repeticiones. El resultado fueron 3542 tRNAs, distribuidos entre los aminoácidos y los tipos de célula -u organelo- de la forma indicada en el Cuadro 7.2.

El alineamiento de las secuencias en la base de datos sigue la numeración estándar, que puede verse en la Figura 7.3, y que va desde la posición 1, en el extremo $5'$, hasta la 73, en el extremo $3'$, seguidas por el **cca** que se liga al aminoácido. A éstas se agregan algunas bases "optativas", que con frecuencia están ausentes: una después de la 17 (17:1), dos después de la 20 (20:1 y 20:2), y las bases de "loop suplementario" que muchos tRNAs presentan entre el loop del anticodón y el loop TΨC, numeradas 47:1, … , 47:16, según sea el largo del loop; además, la base 17, pese a su numeración fija, también puede estar ausente. Los cuatro "tallos" están formados por los pareos de 1–7 con 66–72, 10–13 con 22–25, 27–31 con 39–43 y 49–53 con 61–65, si bien existen excepciones en algunos tRNAs. Los loops van de 14 a 21 (loop D), 54 a 60 (loop TΨC, que en algunos pocos tRNAs está completamente ausente), y 32 a 38 (loop del anticodón, que presenta el anticodón en 34–36). Con eso, las únicas bases libres que no están asociadas a un loop son 8, 9, 26, y 44–48.

Las bases conservadas suelen estar en las partes libres (no pareadas), aunque hay excepciones en ambas direcciones: algunas bases libres no se conservan (como la 59, pese a

GRUPO	*A*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*K*	*L*	*M*	*N*	*P*	*Q*	*R*	*S*	*T*	*V*	*W*	stop	*Y*	*Z**	Total
ARCHAE	18	4	5	7	4	6	4	4	7	15	10	4	9	5	9	11	11	8	5	3	4	1	154
CHLORO	22	12	13	13	9	21	20	29	10	31	18	13	14	9	33	29	23	22	11	8	13	0	373
CY ANI	17	4	4	15	5	10	3	3	12	13	4	5	11	11	7	13	6	13	2	7	7	5	177
CY PLA	3	4	3	1	1	6	0	1	0	3	2	1	2	1	0	12	3	1	2	2	7	0	55
CY SIN	7	2	6	6	8	7	4	8	11	9	6	8	3	14	16	19	11	11	7	4	7	0	174
CYANEL	2	0	0	1	0	1	1	2	0	1	0	0	0	0	0	1	0	0	0	0	0	0	9
EUBACT	59	13	18	20	17	40	16	60	22	56	23	18	27	17	46	52	41	26	17	18	19	4	629
MI ANI	79	78	62	59	70	52	70	72	61	124	45	74	75	60	51	133	63	68	71	32	73	0	1472
MI PLA	2	6	3	6	3	6	9	4	5	4	11	7	8	7	4	15	1	2	7	8	8	0	126
MI SIN	10	12	11	14	10	17	13	13	13	24	26	15	16	17	18	18	13	16	16	10	18	0	320
VIRUS	1	2	1	1	1	5	1	3	4	4	2	3	2	5	2	3	5	1	4	1	2	0	53
Total	220	137	126	143	128	171	141	199	145	284	147	148	167	146	186	306	177	168	142	93	158	10	3542

Cuadro 7.2: Cantidad de tRNAs extractados de la base de datos en [Sprinzl et al, 1998], por grupo y por aminoácido. La clasificación de grupos que los autores usan es: VIRUS(virus o fago), ARCHAE (arqueobacterias), EUBACT (eubacterias), CYANEL (*cyanelle*, organelo fotosintético, ignoro la traducción), CHLORO (cloroplasto), MI SIN (mitocondria de unicelular u hongo), MI PLA (mitocondria vegetal), MI ANI (mitocondria animal), CY SIN (citoplasma de unicelular u hongo), CY PLA (citoplasma vegetal), CY ANI (citoplasma animal). *: *Z* es selenocisteína, el "aminoácido 21", codificado en algunas -escasas- especies, por un codón de término *y además* por la presencia de cierto patrón en la secuencia que lo rodea.

estar rodeada de bases muy conservadas), y algunas bases pareadas en los extremos de los tallos sí lo hacen. La información sobre las bases conservadas varía de acuerdo a la fuente consultada, dependiendo de la información que los autores hayan usado y del porcentaje de ocurrencia que hayan exigido para considerar conservada a una base. Nuestras fuentes fueron un autorizado libro de texto, [Lewin, 2002], otro texto sobre edición de RNA que básicamente repite las posiciones dadas por Lewin con algunas omisiones [Bass, 2001], y otro libro de texto, un poco más antiguo, que agrega varias posiciones conservadas más [Singer y Berg, 1992], al parecer incluyendo algunas con porcentajes relativamente bajos de apariencia, a juzgar por nuestros cálculos (para la base 17, por ejemplo, da **u**, que sólo ocupa esa posición en el 49 % de nuestros tRNAs).

Para confirmar o completar esta información, calculamos las bases conservadas en los tRNA de la base de datos, exigiendo distintos porcentajes. Los resultados para algunos porcentajes se muestran en el Cuadro 7.3, junto con los de las fuentes consultadas, para todas las posiciones libres (loops y "articulaciones") y para las posiciones más conservadas de los tallos. Hemos eliminado las bases no conservadas del loop D (17, 17:1, 20:1, 20:2), y una base libre solitaria (48). Para el resumen de la columna derecha, hemos agregado entre paréntesis los resultados de nuestros propios cálculos: En primer lugar, se agregaron las bases conservadas al nivel de un 75 % en la base de datos (7, 22, 52, 62), dejando fuera sólo algunas aisladas en los tallos. En segundo lugar, se agregaron dos bases de los loops (38 y 59) que no se listan usualmente como conservadas, pero muestran una tendencia clara en la base de datos (71 % y 58 %, respectivamente). [La ausencia de la posición 13 puede parecer rara, pues su base pareada, 22, está incluida (y un porcentaje alto en una implicaría uno alto en la otra). Sin embargo, un 32.5 % de los tRNAs **no** parean sus bases en esa posición, lo que explica la menor ocurrencia de pirimidinas.]

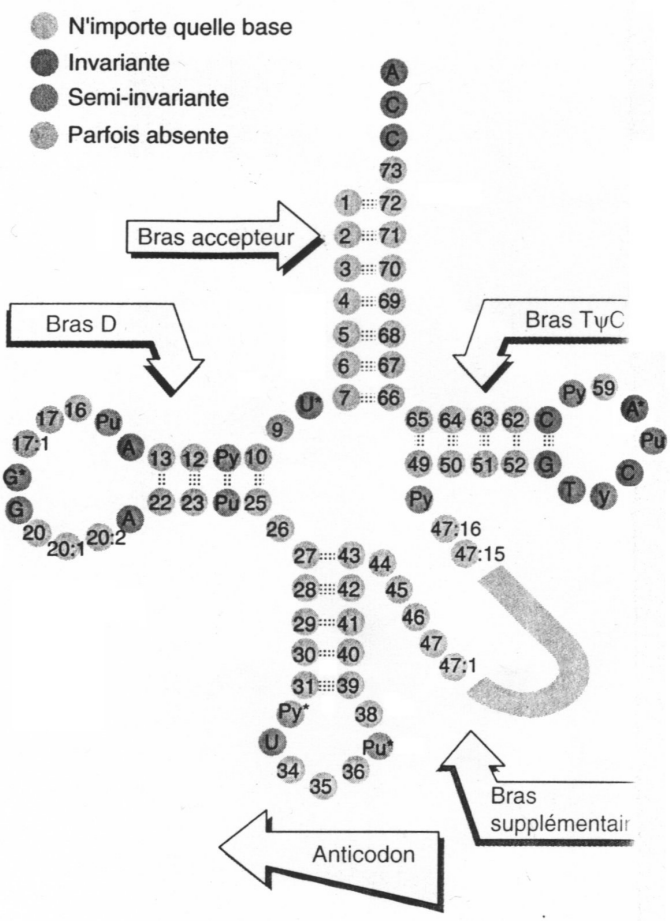


Figura 7.3: Esquema general del tRNA, en la forma presentada por [Lewin, 2002] (edición francesa, Figura 5.3). Los asteriscos indican bases modificadas durante la síntesis del tRNA; Pu y Py representan purinas y pirimidinas, respectivamente (R e Y en nuestra notación).

Nota 1: Tal como se ve en la Figura 7.3, parte de las bases conservadas del tRNA son bases modificadas; además, aparecen tanto **u** como **t**. Hemos convertido todas a su base original (a partir de la cual se ha creado la base modificada), para mantenernos en el alfabeto de cuatro letras, y no nos ocuparemos del tema de las modificaciones de bases, pese a que estas son importantes para el funcionamiento contemporáneo del tRNA, y no se puede descartar del todo que también lo hayan sido en el mundo primitivo.

Nota 2: Interesa saber cuales posiciones conservan el tipo de base, aún cuando no conserven la base específica (es decir, preservan R ó Y). Esto, por dos razones: por un lado, es más fácil un reemplazo de una purina (pirimidina) por otra purina (pirimidina) a lo largo de la evolución (en la jerga del campo, *transiciones* son más frecuentes que *transversiones*). Por otro lado, entre una purina y una pirimidina pueden haber pareos no estándar (como **g-u**), menos estables que los de Watson-Crick, pero aún así útiles para la estructura.

	Posición	Lewin	Bass	Singer	0.55 %	0.6 %	0.65 %	0.7 %	0.75 %	0.8 %	0.85 %	0.9 %	Resumen
tallo aceptor	7				R	R	R	R	R				(R)
libres	8	**u**	**u**	**u**	**u**	**u**	**u**	**u**	**u**	**u**	**u**		**u**
	9			R	**a**	**a**	R	R	R	R	R	R	R
tallo	10			**g**	**g**	**g**	**g**	**g**	**g**	**g**	R	R	**g**
loop D	11	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y
	14	**a**	**a**	**a**	**a**	**a**	**a**	**a**	**a**	**a**	**a**	R	**a**
	15	R	R	R	R	R	R	R	R	R	R		R
	16			Y	Y	Y	Y						Y
loop	18	**g**	**g**	**g**	**g**	**g**	**g**	R	R				**g**
D	19	**g**	**g**	**g**	**g**	**g**	**g**	R					**g**
	20			**u**	Y	Y							**u**
	21	**a**	**a**	**a**	**a**	**a**	**a**	**a**	**a**	**a**	R		**a**
	22				R	R	R	R	R	R			(R)
tallo	24	R		R	**g**	R	R	R	R	R	R		R
loop D	25			**c**	**c**	**c**	**c**	Y	Y	Y	Y	Y	**c**
libre	26				R	R	R	R	R	R	R		(R)
	32	Y	Y	Y	**c**	Y	Y	Y	Y	Y	Y	Y	Y
loop	33	**u**	**u**	**u**	**u**	**u**	**u**	**u**	**u**	**u**	**u**	**u**	**u**
del	34–36							*anticodón*					
anticodón	37	R		R	**a**	**a**	**a**	**a**	**a**	R	R	R	R
	38				**a**	**a**	**a**	R					(R)
tallo	52				**g**	**g**	**g**	**g**	R	R	R		(R)
loop TΨC	53	**g**	**g**	**g**	**g**	**g**	**g**	**g**	**g**	**g**	R		**g**
	54	**u**	**u**	**u**	**u**	**u**	**u**	**u**	**u**	**u**	Y		**u**
	55	**u**	**u**	**u**	**u**	**u**	**u**	**u**	**u**	**u**	Y		**u**
loop	56	**c**	**c**	**c**	**c**	**c**	**c**	**c**	Y	Y			**c**
TΨC	57	R	R	R	R	R	R	R	R	R	R		R
	58	**a**	**a**	**a**	**a**	**a**	**a**	**a**	**a**	**a**	R		**a**
	59				R								(R)
	60	Y	Y	Y	**u**	**u**	Y	Y	Y				Y
tallo	61	**c**	**c**	**c**	**c**	**c**	**c**	**c**	**c**	**c**	Y	Y	**c**
loop TΨC	62				**c**	**c**	**c**	Y	Y	Y	Y		(Y)

Cuadro 7.3: Bases conservadas según distintas fuentes: [Lewin, 2002], [Bass, 2001], [Singer y Berg, 1992], y nuestros propios cálculos con distintos niveles de ocurrencia a partir de la base de datos de [Sprinzl et al, 1998].

7.3. Alineamiento de anillos y tRNA

Nos interesaba a continuación comparar anillos con las partes libres y/o conservadas de los tRNAs. Ya que estas partes no son contiguas, debían considerarse distintas formas de cortar cada anillo, y alinear las partes con las partes conservadas de los tRNAs en la base de datos, en los cinco segmentos contiguos en que éstas se distribuyen.

Lo que se hizo fue, para cada anillo de A_0 (es decir, para la secuencia listada y sus 21 rotaciones), probar todas las formas de cortarlo en a lo más 4 segmentos, cada uno de tamaño no menor a 3 bases, y para cada uno de estos cortes alinear los segmentos (en el orden $5' \to 3'$) con los segmentos contiguos del Cuadro 7.3 (7–11, 14–22, 24–26, 32–38 y 52–62), exigiendo que al menos las partes más conservadas de los loops (14–21, 32–37 y 54–58) quedaran cubiertas. Esto dio un total de 26 esquemas a comparar, para cada rotación de cada anillo, con cada secuencia de tRNA; los 26 esquemas se muestran en el Cuadro 7.4.

En cada comparación se calcularon dos "puntajes": uno por bases idénticas, y otro que asignaba un punto a bases idénticas, y medio punto a bases del mismo tipo (puri-

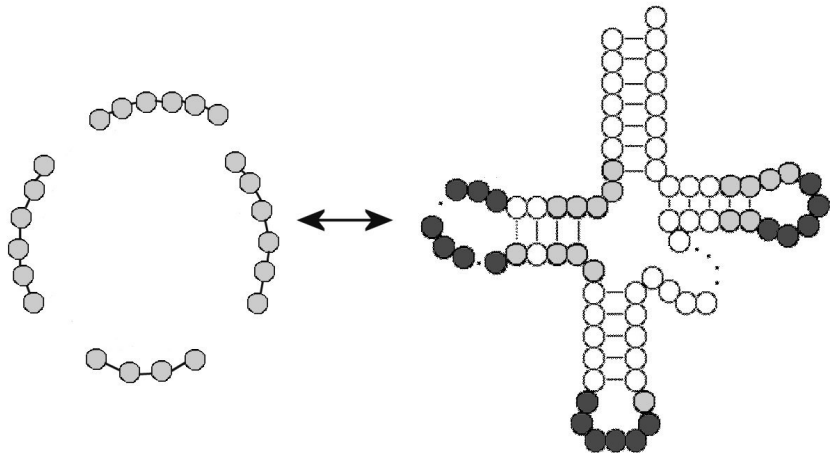


Figura 7.4: Alineamiento: los anillos se cortaron en a lo más cuatro segmentos, de no menos de tres bases. Para todas las combinaciones posibles (pero manteniendo el orden $5' \to 3'$) se hizo el alineamiento con los segmentos en gris de todos los tRNA, exigiendo que las posiciones en negro estuviesen cubiertas.

$0:7:0:6:9$	$0:7:0:7:8^{+}$	$0:8:0:6:8^{++}$	$3:7:0:6:6^{+}$	$3:7:0:7:5^{++}$	$4:7:0:6:5^{++}$	$0:7:3:7:5^{++}$
$0:7:0:6:9^{+}$	$0:7:0:7:8^{++}$	$0:8:0:7:7$	$3^{+}:7:0:6:6^{+}$	$3^{+}:7:0:7:5^{++}$	$4^{+}:7:0:6:5^{++}$	$0:8:3:6:5^{++}$
$0:7:0:6:9^{++}$	$0:8:0:6:8$	$0:8:0:7:7^{+}$	$3:7:0:6:6^{++}$	$3:8:0:6:5^{++}$	$0:7:3:6:6^{+}$	
$0:7:0:7:8$	$0:8:0:6:8^{+}$	$0:8:0:7:7^{++}$	$3^{+}:7:0:6:6^{++}$	$3^{+}:8:0:6:5^{++}$	$0:7:3:6:6^{++}$	

Cuadro 7.4: Esquemas para corte y alineamiento: la entrada $0:8:0:6:8^{++}$, por ejemplo, indica el corte de la secuencia (del anillo) en trozos de 8, 6 y 8 bases, que se alinearán con el segundo, cuarto y quinto segmento conservado del tRNA (en orden $5' \to 3'$), partiendo de la primera base de cada segmento, salvo el último, que se ubicará dos posiciones más adelante (es decir, desde la posición 54).

nas/pirimidinas). A continuación, los anillos fueron rankeados de acuerdo al mejor puntaje que hubiesen obtenido (para algún corte y algún alineamiento) en relación a algún tRNA en particular y también en promedio con la base de datos completa. Para no dejar dudas, lo que se define es:

$$
\begin{aligned}
E(x) &= \max_{0\le i<22}\ \max_{0\le j<26}\ \max_{1\le k\le 3542}\ m(\sigma^i(x), s_j, t_k) \\
E^*(x) &= \max_{0\le i<22}\ \max_{0\le j<26}\ \max_{1\le k\le 3542}\ \frac{1}{2}\left(m_{RY}(\sigma^i(x), s_j, t_k) + m(\sigma^i(x), s_j, t_k)\right) \\
pE(x) &= \max_{0\le i<22}\ \max_{0\le j<26}\ \frac{1}{3542}\sum_{1\le k\le 3542} m(\sigma^i(x), s_j, t_k) \\
pE^*(x) &= \max_{0\le i<22}\ \max_{0\le j<26}\ \frac{1}{3542}\sum_{1\le k\le 3542}\ \frac{1}{2}\left(m_{RY}(\sigma^i(x), s_j, t_k) + m(\sigma^i(x), s_j, t_k)\right)
\end{aligned}
$$

donde $m(\sigma^i(x), s_j, t_k)$ es la cantidad de bases coincidentes al alinear $\sigma^i(x)$ con el k-ésimo tRNA según el j-ésimo esquema, y m_{RY} es lo mismo, pero exigiendo sólo que las bases sean ambas purinas o ambas pirimidinas (de modo que $\frac{m_{RY}+m}{2}$ suma 1 por cada base

idéntica –que se cuenta en ambas– y $\frac{1}{2}$ por las que sólo son del mismo tipo).

	E		pE		E^*		pE^*	
Posición	Anillo	Valor	Anillo	Valor	Anillo	Valor	Anillo	Valor
1	20766	20	20766	12.93	20766	21	20766	15.09
2	20782	20	8585	12.69	20764	20.5	8556	14.79
3	10938	19	10938	12.52	14818	20.5	5970	14.77
4	13930	19	13930	12.47	20782	20.5	20764	14.76
5	14742	19	9264	12.42	14034	20.5	13930	14.62
6	23824	19	14742	12.34	13921	20.5	20743	14.58
7	8549	19	23824	12.22	14011	20.5	14742	14.55
8	14270	19	8549	12.18	14736	20.5	20786	14.54
9	14582	19	14270	12.17	11007	20.5	14818	14.50
10	23586	19	8544	12.15	8556	20	19388	14.46
11	20764	19	14582	12.11	5970	20	8540	14.45
12	14890	19	23586	12.09	13930	20	5968	14.43
13	7041	19	20764	12.06	20743	20	8549	14.41
14	14748	19	8556	12.05	14742	20	20782	14.41
15	5099	19	27276	12.04	20786	20	8585	14.35
16	8734	19	14890	12.04	8540	20	14034	14.33
17	27292	19	5970	12.01	5968	20	22354	14.32
18	14922	19	7041	12.01	8549	20	20731	14.32
19	20786	19	14748	11.99	22354	20	6142	14.31
20	23787	19	5099	11.96	14890	20	14890	14.30
21	23408	19	20731	11.95	14846	20	14846	14.29
22	10836	19	6701	11.94	22242	20	22242	14.29
23	10840	19	5102	11.90	23586	20	13921	14.28
24	21366	19	10909	11.89	8555	20	8036	14.27
25	28443	19	3672	11.89	9000	20	10909	14.27

Cuadro 7.5: Anillos mejor ranqueados al alinear A_0 con la base de datos de tRNA.

Para cada una de estas cantidades (E, E^*, pE y pE^*), se rankearon todos los anillos de A_0; los primeros 25 anillos de cada ranking se muestran en el Cuadro 7.5. Llama la atención, en primer lugar, que existieron 2 anillos con un alineamiento de *20 bases idénticas* con algún tRNA. Además, el máximo en E^* es 21, por lo que existe un anillo que al alinearse con un tRNA logra 20 bases idénticas y dos bases distintas pero del mismo tipo (es decir, el alineamiento ocupa las 22 bases, pero con dos "transiciones").

Más notable aún es el comienzo del Cuadro 7.5; en él encontramos, en las cuatro columnas, al mismo anillo, 20766, ocupando el primer lugar. Lo notable no es sólo que el mismo anillo se destaque en los cuatro valores, sino que *se trata del anillo 13 de A^*, ¡el anillo "ganador" en la sección 6.3!*. El sólo hecho de que aparezca un anillo con un nivel tan alto de alineamiento no es tan sorprendente, pues el histograma de los valores de, por ejemplo, pE, es similar al comparar A_0 y un conjunto de secuencias aleatorias con la misma frecuencia de bases (ver Figura 7.5); lo realmente notable es que el ganador sea el mismo anillo que se destacó antes por razones completamente independientes.

A este anillo, que ha resultado ser el número 1 en A_0 por dos vías completamente dis-

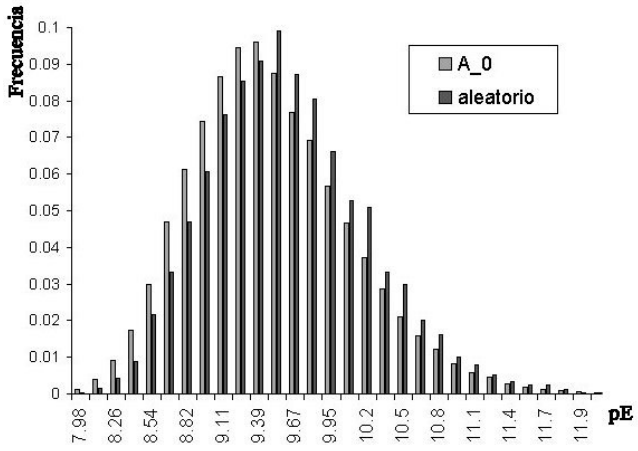


Figura 7.5: Histograma de pE en A_0 y en un conjunto de secuencias aleatorias con igual frecuencia de ocurrencia de bases.

tintas, lo hemos llamado anillo AL (por *Archetypical Loops ring*), y ocupa ahora, en la argumentación completa, el lugar que ocupaba el anillo AB en [Demongeot y Besson, 1996]. Sin embargo, tiene varias diferencias respecto de AB: en primer lugar, no ha sido escogido *por estar cercano a los tRNA*, sino que ha *resultado estarlo*, una vez exigidas la condición de hairpin óptimo (mucho mejor que el de AB) y de menor distancia de los demás que tenían hairpins tan buenos como el suyo. Además, su alineamiento es mejor: en lugar de cortarlo en 5 trozos, sólo lo cortamos en 4, y ninguno de ellos queda invertido (como era el caso en AB) sino que se respeta el orden $5' \to 3'$.

	Primera articulación			Loop D							Loop del anticodón					Loop TΨC						
Posición	7	8	9	14	15	16	18	19	20	21	32	33	34–36	37	38	54	55	56	57	58	59	60
Conservadas	(R)	**u**	R	**a**	R	Y	**g**	**g**	**u**	**a**	Y	**u**	anticodón	R	(R)	**u**	**u**	**c**	R	**a**	(R)	Y
Anillo AL	**a**	**u**	**g**	**a**	**a**	**u**	**g**	**g**	**u**	**a**	**c**	**u**	gcc	**a**		**u**	**u**	**c**	**a**	**a**	**g**	
Marchantia polym.	**a**	**u**	**a**	**a**	**a**	**u**	**g**	**g**	**u**	**a**	**u**	**u**	gcc	**a**	**a**	**u**	**u**	**c**	**a**	**a**	**g**	**u**

Cuadro 7.6: Alineamiento del anillo AL con la secuencia de consenso y con el Gly-tRNA mitocondrial de *Marchantia polym.*

El esquema en que AL dio su alineamiento óptimo es $3{:}7{:}0{:}6{:}6^{++}$ con la rotación **augaauggguacugccauucaag**; es decir, los segmentos **aug**, **aauggua**, **cugcca** y **uucaag** se ubican en las posiciones 7–9, 14–21, 32–38 y 54–59, respectivamente. El tRNA con el que se da el mejor alineamiento es un Gly-tRNA mitocondrial de *Marchantia polym.*, un vegetal (de hecho, pariente cercano de *Œnothera*). En el Cuadro 7.6 se muestra el alineamiento; allí puede verse que además AL satisface la secuencia de consenso (bases conservadas) en *todas* las posiciones. Un poco más: si vemos el Cuadro 7.3, podemos notar que las bases usadas por AL en las posiciones 32, 37, 38 y 60 coinciden con el 55, 75, 65 y 60 por ciento de los tRNA de la base de datos; la única excepción es la posición 9, en que si bien AL satisface la existencia de una purina, ésta es **g** para AL y **a** en el 60 % de la base de datos.

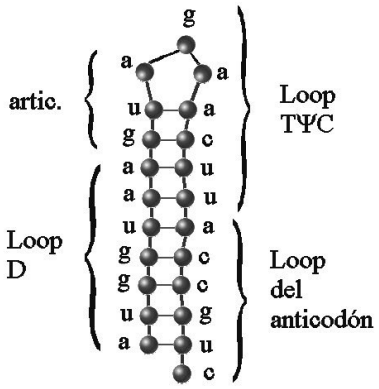


Figura 7.6: El hairpin de AL separa de manera natural los segmentos de alineación; el punto en que se abre, y el punto de ruptura más probable del hairpin (en el loop) producen dos de los cuatro cortes.

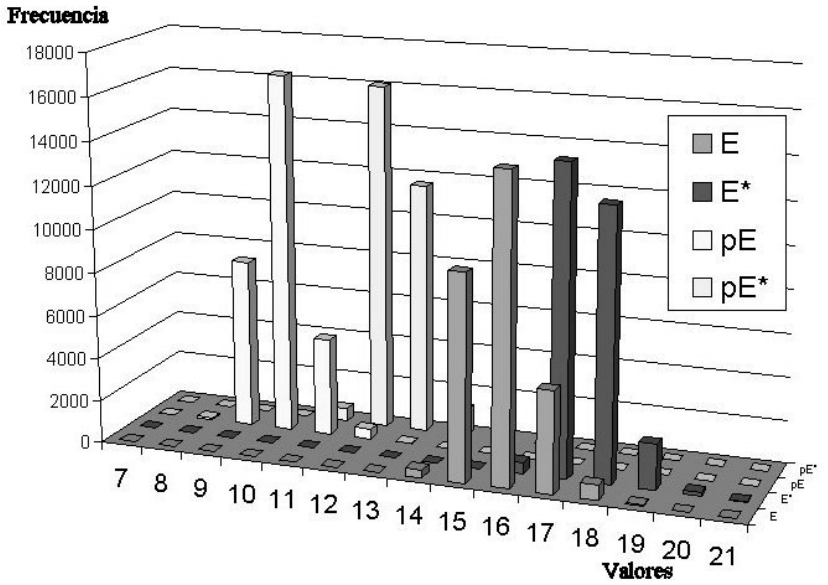


Figura 7.7: Histograma de los valores E, E^*, pE y pE^* en A_0.

Cabe hacer notar que otros dos anillos del cluster de AL en A^*, que ocupaban el segundo y cuarto lugar en el ranking del Cuadro 6.13, aparecen también en el Cuadro 7.5: son el 5 (numerado aquí 20786) y el 7 (numerado 20731), y figuran en 3 y 2 columnas, respectivamente.

Un anillo de interés es el 20782, que aparece igualado a AL en la columna de E (mejor matching exacto con un tRNA), y también figura en la columna de pE^*; podría haber figurado en la de E^*, pues también alcanza el valor 20, mientras que pE lo sitúa en el lugar 114 dentro de A_0. Su alineamiento óptimo es con un Lys-tRNA del cloroplasto de

Nicotiana tabacum (otra vez, una planta, y no lejana a *Marchantia* y *Œnothera*). Este anillo no entró en A^* por tener un hairpin bastante malo (de largo 6); se trata de un anillo bastante similar a AL, pues comparten una subcadena de largo 17. De hecho, al evaluar d_H^c con respecto a AL en todos los anillos de A_0, el 20782 está dentro del 0.07 % con menor distancia.

Esta última evaluación (la distancia a AL en todo A_0) permitió notar que existe una fuerte correlación entre la distancia a AL y la ubicación de los anillos en el ranking 7.5; esto es razonable, pues AL satisface la secuencia de consenso, que por definición es la mejor aproximación al conjunto completo de los tRNA. Parecerse a AL, por lo tanto, significa también estar cerca de estos últimos. En todo caso, sería interesante poder usar nuestras medidas de distancias para estudiar la estructura de clusters, no ya del pequeño conjunto A^*, sino de todo A_0; esto es computacionalmente pesado, pero debería ser factible.

Una última observación relevante es respecto al hecho de que el tRNA más cercano a AL haya sido un Gly-tRNA, o, visto desde el otro lado, el hecho de que AL coloque **gcc** en la posición del anticodón. La glicina es el aminoácido más simple, y el de formación más común en los experimentos de síntesis abiótica. Por esto, y basándose a veces también en intentos de filogenia del tRNA, se ha sugerido que el primer tRNA pudo haber estado asociado a este aminoácido. Más aún: ya desde Eigen [Eigen et al., 1981] se ha propuesto a los codones (complementarios) **gcc** y **ggc** como los codones más antiguos, con **ggc** (cuyo anticodón es **gcc**) codificando a la glicina. Sea cual sea la historia que se construya desde el escenario de los anillos hacia el mundo posterior, un dato relevante es que AL, al alinearse con los tRNA (y de manera óptima con Gly-tRNAs) pone precisamente **gcc** en el anticodón.

Capítulo 8

Síntesis y conclusiones

El propósito de este trabajo ha sido revisitar la idea propuesta por J. Demongeot (y presentada en el Capítulo 4) sobre una cadena circular de RNA que podría haber jugado un rol clave en el inicio de la vida. Se ha agregado información, cálculos, puntos de vista, algunas ideas biológicas y otras combinatorias, y se ha logrado no sólo reducir las hipótesis, sino también encontrar una serie de resultados nuevos (como una posible explicación para el surgimiento de la forma actual de los genes), y una argumentación que conduce de manera completamente independiente a una única secuencia, que resulta ser coincidente con las bases conservadas de los tRNA actuales (a diferencia de la secuencia AB en [Demongeot y Besson, 1996], que había sido escogida precisamente *en función* de su semejanza con un tRNA moderno particular). A continuación, un resumen de los principales puntos de los capítulos anteriores, de acuerdo a la forma en que se insertan en el marco de la teoría de Demongeot.

8.1. Resumen: los anillos bajo nueva luz

El cuadro inicial

La idea de Demongeot parte planteándose un escenario prebiótico, en el que bases (en rigor, nucleótidos) y aminoácidos flotan libremente o en pequeñas cadenas, cuya formación pudo haberse visto facilitada, por ejemplo, por la actividad catalizadora de una arcilla, como se ha propuesto en varios modelos del origen de la vida. Adoptando una versión fuerte de la teoría estereoquímica del código genético, señala que las cadenas de nucleótidos, esto es, pequeños RNA, pudieron verse estabilizadas (y por lo tanto, haber aumentado en abundancia) al asociarse con aminoácidos de acuerdo a una afinidad entre últimos y los codones que los codifican. En esas circunstancias, sugiere que la "mejor estrategia" para un RNA primitivo habría sido codificar simultáneamente todos los aminoácidos presentes en el código, y hacerlo en forma circular (un anillo), de longitud mínima. Al no haber anillos de largo 20 que codifiquen los 20 aminoácidos, agrega una restricción adicional:

que los anillos tengan la forma de los genes modernos, es decir, que se puedan escribir como (**aug**, … , stop), repitiendo por lo tanto el codón de la metionina (que oficia de *start* en los genes modernos) y codificando además un stop. Con esto, los anillos son de largo 22, y encuentra 1280 soluciones posibles (el conjunto que aquí hemos llamado A_1), de las cuales una, el anillo AB, es destacada por ser la más parecida a los loops de un Gly-tRNA de *Œnothera*.

¿Por qué un stop, y largo 22?

Nuestras simulaciones han indicado que codificar una vez cada uno de los aminoácidos presentes en el medio parece efectivamente ser la mejor estrategia para ganar el "juego" descrito; sin embargo, el paso siguiente, de exigir la forma (**aug**, … , stop), resulta en extremo arbitrario, pues en el escenario prebiótico no hay mecanismo alguno para el que esa forma pueda significar algo. Nuestra generación de listados de anillos de distintos largos, codificando distintos aminoácidos, muestran que efectivamente es imposible codificar 20 con 20 bases (y sugieren también lo curioso de esa imposibilidad, pues para longitudes menores la combinatoria del código se muestra bastante más flexible), pero muestran además que existen 3456 anillos de largo 21 que codifican 20 aminoácidos, repitiendo uno (siempre, triptofano). ¿Por qué no pudieron ser esos los anillos primitivos?

Aquí otra observación hecha en las simulaciones da una pista: al eliminarse uno de los aminoácidos del medio (virtual), la ventaja selectiva del anillo que codificaba todas las clases *aumentó*. Al parecer, una estructura demasiado estable se reproduce con menos frecuencia (ya que aquí la reproducción se da por ruptura casual, y subsiguiente pareo con bases libres o cadenas más cortas); al ser fija (dentro del anillo) la posición de ruptura, se habría además asegurado la presencia de cadenas libres con la misma secuencia. Aún falta estudiar más en detalle esta característica del juego, pero los resultados de la simulación estarían indicando que para los anillos primitivos pudo ser una buena "opción" (es decir, una característica que de estar presente, ayudaba a sobrevivir o, en este caso a reproducirse) el codificar, además de los 20 aminoácidos, un "stop", que en este punto no habría tenido el significado de stop que tiene en la maquinaria actual, sino que simplemente habría sido un punto sistemáticamente débil del anillo, al no ser afín a ninguno de los aminoácidos.

Repetición "espontánea" de *aug*

Al pedir codificar las 21 clases de equivalencias del código en longitud mínima, encontramos que el largo mínimo de los anillos es 22, y que las soluciones (que aquí hemos llamado A_0) son 29520. Y aquí viene una primera sorpresa: el codón que "sobra" (pues en 22 estamos pidiendo sólo 21 clases) es siempre de la forma **auN** (con **N** representando cualquier base), y en el **52 %** de los anillos, esa **N** es **g**. En otras palabras, la condición impone la repetición mayoritaria del codón que codifica *start*, sugiriendo que este recibió ese significado precisamente por haber estado "disponible"; más abajo se retomará esto.

Cercanía de A_0 con los genomas actuales

Otras características interesantes de los anillos de A_0, que incluyen a los de Demongeot y

los generalizan (al no imponer una forma de gen, sino sólo la codificación de las 21 clases), son en primer lugar que (como ya había observado Demongeot en A_1) ninguno de ellos incluye el dinucleótido **cg**, que es conocido por ser el más raro en la mayor parte de los seres vivos (con excepciones en las bacterias), mientras que en cambio, *todos* los demás dinucleótidos están presentes. En segundo lugar, al calcular el uso relativo de codones (RSCU) en A_0 y en una serie de especies actuales, se encontró correlación positiva en 8 de 12 casos. Si bien es discutible la posibilidad de que las estadísticas de los genomas actuales reflejen aún algunos rasgos de los genomas primitivos, tras billones de años de evolución, en caso de que se admita tal posibilidad estos resultados sobre A_0 son argumentos que se suman a la idea de que este conjunto pudo estar presente en el inicio de la vida.

Codificación en anillos complementarios

También resultó interesante observar que A_0, en comparación con anillos escogidos al azar, tiene un fuerte sesgo a favor de la codificación de un número de aminoácidos alto en los anillos *complementarios* a los de A_0; la importancia de esto es que la reproducción de los anillos pasa por sus complementos, y por lo tanto es "útil" que estos, o sus subcadenas, tengan también la "buena propiedad" de afinidad con aminoácidos distintos. Evidentemente, el óptimo sería que el complemento *también* codificara los 20 aminoácidos, pero esto no es posible en largo 22, sino recién para anillos de largo 24. ¿Por qué no fueron estos los anillos "elegidos"? Pudieron serlo; sin embargo, también hay argumentos en su contra: el excedente de codones para algunos aminoácidos pudo haberlos hecho menos estables, y también su longitud mayor pudo haber conspirado en su contra; además, es probable que la reproducción no fuese principalmente de anillos completos, sino de subcadenas, y en ese caso lo importante no es que el anillo complementario codifique todas las clases, sino que la mayoría de sus subcadenas codifiquen sólo clases distintas, para lo cual basta con el sesgo en A_0.

Hairpins, y una nueva selección

Demongeot había observado que el anillo AB podía adoptar una estructura de *hairpin*; esto nos sugirió estudiar las estructuras de hairpin en A_0, y más en general, su grado de palindromía (distancia respecto al complemento). Pudimos observar en primer lugar que la frecuencia de ocurrencia de buenos hairpins, y de distancias pequeñas respecto del complemento, son algo mayores en A_0 que en una secuencia aleatoria, aunque se aproximan bastante a lo que ocurren en secuencias aleatorias que comparten la frecuencia de ocurrencia de las bases de A_0.

Llegados a este punto, agregamos un elemento al "juego" del mundo primigenio: bajo condiciones ambientales adversas, debió ser una gran ventaja selectiva para una secuencia el poder adoptar la forma más estable, "encapsulada", de un hairpin. Con eso en mente, calculamos los mejores hairpins para todos los anillos, y escogimos un subconjunto, A^*, formado por las 24 secuencias de A_0 capaces de formar un hairpin con un tallo de largo máximo (9, para cadenas de largo 22) y loop mínimo (3).

Este conjunto, denotado A^* y del cual el antiguo anillo AB de Demongeot queda excluido, presenta varias características notables. El uso relativo de codones, por ejemplo, resulta

ser representativo del que muestra todo A_0, y por lo tanto, también se correlaciona positivamente con los genomas contemporáneos que se usaron como punto de comparación. Más interesante aún es el hecho de que en A^* no ya el 52 %, sino el **79 %** de los anillos repite el codón **aug**, esto es, la metionina (M). No sólo eso: la forma (**aug**, ... , stop), de los genes modernos, la presentan el 38 % de los anillos de A^*. Si recordamos que en [Demongeot y Besson, 1996] esta forma se había impuesto a las secuencias, resulta sorprendente que ahora *emerja de consideraciones completamente independientes*.

Emergencia de la forma *aug*... *stop*

¿Qué puede significar esto? La respuesta no la conocemos, pues en el escenario que discutimos, la maquinaria de la traducción aún está ausente; sin embargo, en algún momento esa maquinaria debió empezar a evolucionar a partir de estas mismas estructuras. Sin adentrarnos mucho en la forma en que eso pudo ocurrir -un tema pendiente, que escapa a nuestro trabajo y también a nuestra área-, podemos imaginar que si las primeras cadenas disponibles eran los propios anillos, con su stop "natural", y la mayoría de estos anillos se rompían en cadenas de la forma (**aug**, ... , stop), la incipiente maquinaria debió ajustarse a copiar cadenas de esa forma; la posterior evolución de genes más largos pudo darse por inserciones y repeticiones, pero la forma del comienzo y final ya habría quedado fijada.

Un juego adicional: parecerse a los demás

Volviendo al juego primitivo: nos encontramos con este interesante conjunto de anillos, A^*, con las buenas propiedades de A_0 y además poseedores de la ventaja selectiva de un excelente hairpin (que va asociado además, de manera natural, a una alta palindromía, lo cual también es positivo en un juego de cadenas que se rompen y rearman). Dentro de esos 24 anillos, ¿quién ganaba? Nuestra idea es que, en igualdad de condiciones, y suponiendo otra "etapa" del juego, determinada por la ruptura y rearmado de las cadenas, los anillos a la larga más abundantes serían aquellos que se parecieran más, en promedio, a los demás que están presentes. Este proceso, que algunas simulaciones preliminares parecen confirmar, tendría además una retroalimentación positiva, pues al verse favorecido aquel que se parezca más a los demás, aumentará su abundancia, y con ello su ventaja.

¿Cómo precisar la idea de "más parecido"? Para esto construimos varias medidas de distancia entre anillos, intentando capturar las características del juego al que habrían estado sometidos. El resultado fueron dos métricas y una semi-métrica, que se corresponden aproximadamente con los mecanismos evolutivos de mutación, transposición y reordenamiento que tienen lugar en el material genético. Al evaluar estas distancias en A^*, y promediar, aparecen claramente dos anillos (el número 7, y en especial el número 13 del Cuadro 6.9) con distancias menores hacia los demás. Un análisis más fino, mediante técnicas de *clustering*, permitió ver la estructura de A^* en función de las distintas métricas, distinguiéndose la presencia de un grupo dominante de anillos, encabezados por el número 13.

Comparación con tRNAs actuales: el anillo AL

El tRNA es el fósil molecular por excelencia, tanto por su función como por su universalidad. Y su parte fósil son las posiciones conservadas, ubicadas básicamente en las partes

libres (principalmente los loops) y en los extremos de algunos tallos. Luego de un intento infructuoso de recopilar información a partir de Genbank, recurrimos a la base de datos de [Sprinzl et al., 1998], de la que obtuvimos las secuencias alineadas de 3542 tRNAs; esto nos permitió, en lugar de guiarnos por un tRNA en particular, como en [Demongeot y Besson, 1996], trabajar con un conjunto amplio proveniente de todos los grupos principales de seres vivos, y contrastarlos todos contra todos los anillos de A_0.

El resultado es sorprendente: al hacer el alineamiento con los tRNA, el que se destaca, tanto en valores promedios como en el alineamiento máximo con un tRNA particular, es precisamente el anillo 13, destacado en A^* por razones completamente independientes. En particular este anillo, rebautizado AL (*Archetypical Loop ring*), cumple con la secuencia de consenso de los tRNA, y se diferencia de las partes libres del Gly-tRNA mitocondrial de *Marchantia* sólo por dos transiciones (cambios de purina por purina, o pirimidina por pirimidina). Entre otras características de interés (aparte de las que lo señalaron en A^* y las del alineamiento): su hairpin separa claramente los distintos dominios de alineamiento, y pone **gcc** (propuesto por Eigen como uno de los anticodones originarios) en la posición del anticodón.

La ubicación del alineamiento de AL en el tRNA merece un comentario. Básicamente, se alínea con los loops: el loop D completo, buena parte del loop TΨC (exceptuándose la posición 60, en general menos conservada y que no participa en interacciones tridimensionales), y también casi todo el loop del anticodón, exceptuando la posición 38. Respecto a esto último resulta interesante que en los tRNAs con intrones del grupo I (probablemente los más antiguos intrones), el intrón está insertado precisamente entre las posiciones 37 y 38, lo que sugiere, en caso de que los intrones hayan estado presentes ya en el momento de la formación de los primeros tRNA, que la base 38 pudo simplemente no haber sido contigua a la 37.

Finalmente, hay un pequeño segmento de AL que en el alineamiento óptimo queda situado en las posiciones 7 a 9. ¿Tiene algún sentido esa ubicación específica? Al buscar una respuesta, nos encontramos con un dato muy significativo, en la estructura tridimensional del tRNA. En esa estructura, el loop D y el del anticodón se sitúan *perpendiculares* al loop TΨC, produciéndose una serie de interacciones entre las bases de los loops laterales. Y en esa extraña posición, las bases situadas en las posiciones 8 y 9 juegan el rol clave, doblando al loop D "hacia atrás" [Harshey, 2003]. El segmento de AL, por lo tanto, se ubica en un punto crucial para la estructura del tRNA, y su presencia allí debería ser un dato importante a la hora de imaginar el camino desde el mundo de los anillos al del tRNA contemporáneo: es probable que el propio AL haya adoptado ya una estructura tridimensional como la actual.

AL viene a reemplazar al anillo AB de [Demongeot y Besson, 1996] como caballo de batalla de la teoría: aparte de ser más estable (mejor hairpin) y de estar mejor alineado, son también menos arbitrarios los dos caminos (para AB era sólo uno) que lo señalan. En todo caso, esto no implica que sólo AL pueda haber sido relevante en el inicio de la vida: otros anillos de su cluster en A^*, o incluso de A_0, pueden haber jugado un rol también. Mientras no se escriba una propuesta que conecte concretamente el escenario de

los anillos con un mecanismo de traducción primitivo, no será posible juzgar la forma o el número de tipos de anillos que intervinieron en la formación de los primeros tRNAs, y de los primeros genes.

8.2. Contexto actual de la teoría

¿Cómo se sitúa esta teoría en el contexto de las ideas sobre el origen de la vida? Al parecer, llenando un cierto vacío. Existen estudios sobre la química prebiótica, y las circunstancias o mecanismos que habrían hecho posible la polimerización y otras reacciones que requieren de catálisis. Por otro lado, existen estudios sobre las posibles formas primitivas del tRNA, su aminoacilación, etc. Sin embargo, en la zona intermedia entre la química y la primera biología, lo que prima son enfoques que apuntan a los aspectos más teóricos (y no a las secuencias concretas), como los de Eigen, Szathmáry o Kauffman.

La razón principal para esta brecha, a mi entender, es la inexistencia (al menos aparente) de fósiles moleculares que nos permitan ir más allá de los primeros tRNAs, en dirección a las primeras reacciones y las primeras cadenas. Y es ahí donde el escenario propuesto por Demongeot interviene: conecta el mundo prebiótico con los fósiles moleculares más antiguos, las bases conservadas del tRNA, y lo hace (sobre todo en la versión remozada que nace del presente trabajo) a través de consideraciones razonables sobre la forma en que la selección pudo haber actuado sobre las cadenas de RNA en aquella etapa. Desde luego, saltan a la vista muchas preguntas, sobre todo respecto a la forma en que desde este escenario se pudo pasar a los sistemas posteriores. Pero al menos hay una buena propuesta respecto a *lo que estaba flotando allí* cuando esos sistemas aparecieron. A esto se agregan además algunos productos "colaterales" de nuestro trabajo, como la posible explicación para la selección de M como start, y para la forma misma de los genes.

La teoría propuesta se relaciona directamente con varias ideas expuestas en el Capítulo 3. En primer lugar, con la teoría estereoquímica del código genético, de la cual depende fuertemente, y a la que por lo tanto avala gracias a la coincidencia final con los tRNA reales. Existen distintas versiones de la teoría, relacionando a los aminoácidos con los codones, con los anticodones, o con el par codón/anticodón; para efectos del escenario de Demongeot, cualquiera de estas opciones puede funcionar, aunque calza mejor con la afinidad hacia codones. Por desgracia, como hemos visto antes, aún es poca la evidencia experimental en este tema; sin embargo, la poca que hay parece corroborar la posibilidad de la teoría estereoquímica.

Ciertamente, el escenario de Demongeot también podría ser sostenible con versiones más débiles de la teoría estereoquímica. Bastaría, por ejemplo, con que *algunos* codones resultaran ser afines a sus aminoácidos (aunque esos, por supuesto, tendrían que ser los que se están usando en A_0, A^*, o AL). También podría existir una distribución de afinidades, no tan maníquea como la hemos supuesto, pero sí con preferencia de los aminoácidos hacia los codones codificadores. De hecho, el escenario ideal sería que el análisis hecho en esta Tesis pudiese reproducirse a partir, no ya del código genético, sino de una tabla de datos

experimentales respecto a afinidades de codones y aminoácidos.

Otro tema relacionado es la teoría de cuasiespecies de Eigen: desde el momento en que estamos pensando en un conjunto de secuencias, algunas más parecidas que otras, pero con presencia de un núcleo (el cluster central de 6.3) de secuencias muy similares entre sí, nos acercamos al concepto de las cuasiespecies, aunque en una versión un poco más ruda, en que no existe una máquina reproductora con probabilidad de error, sino que la reproducción se da a través de un juego cuya combinatoria determina la evolución del conjunto de cadenas/anillos. Sería sin duda interesante ver el paralelo entre ambas situaciones, y la posible dinámica del paso desde una a la otra.

¿Cómo se pasó de los anillos a los tRNA? ¿Participaron distintos anillos en la formación de distintos tRNA, o de distintas partes del tRNA, o fue un anillo en particular (nuestro candidato es AL) el que "ganó el juego" y dio lugar al primer tRNA, a partir del cual habrían derivado los otros? Las historias imaginables son varias, y encajan de distintas formas con las teorías existentes sobre el tema. Se ha observado, por ejemplo, que existen una cierta complementariedad (en el sentido de Watson-Crick) entre las secuencias de tRNAs asociados a anticodones complementarios. ¿Un anillo y su anti-anillo? Otros han visto en las secuencias del tRNA la evidencia de fusiones de hairpins, en distintas combinaciones, según el autor que se consulte; sería interesante explorar en detalle las relaciones entre este tipo de ideas y el escenario de los anillos.

Parte de la relación con otras teorías, por supuesto, es negativa. El escenario de Demongeot no calza (o se vuelve irrelevante) con las teorías que proponen la arbitrariedad de código genético, o con las teorías sobre un origen de la vida sin polímeros. Sin embargo, con la mayor parte de las ideas del *mainstream* actual parece posible el diálogo, o aún la complementación.

Con todo, si algo se ha obtenido como resultado del presente trabajo, es la certeza de que existe una relación no trivial entre el tRNA y las cadenas minimales que codifican aminoácidos diversos, condimentada con algunos detalles intrigantes como la aparición de la forma del gen. Suficiente como para aseverar que, si se niega nuestra explicación, al menos no puede negarse la presencia de algo que requiere ser explicado, y toda propuesta será bienvenida (es posible, como por ejemplo propuso alguien al final de una charla en que expuse el tema, concebir incluso alguna explicación inversa a la nuestra, que explique el código *a partir* de la existencia de anillos primitivos).

8.3. Caminos a seguir

Hay muchos puntos en este trabajo en los que una investigación mayor podría ser provechosa. En primer lugar están las simulaciones, ninguna de las cuales llegó a ser completamente concluyente; algunas posibles extensiones se discutieron ya en 5.4. Aparte de ellas, también sería bueno lograr una aproximación analítica a los distintos "juegos" planteados, y mejor aún, reunirlos en una sola simulación (en lugar de separar, como aquí se hizo,

la etapa de selección de anillos codificadores minimales, de la etapa de competencia por subcadenas más abundantes, que en realidad son partes simultáneas de un mismo juego).

En cuanto al análisis de los conjuntos de anillos, sería interesante llevar a cabo el estudio de la estructura de clusters en todo A_0; es posible, por ejemplo, que la selección de A^* (por hairpins óptimos) ni siquiera resulte necesaria para llegar a AL. Y en cuanto a los hairpins, sería bueno poder evaluar de manera exacta las energías libres efectivas para todas las secuencias, en lugar de seguir una *rule of thumb* como fue el tomar los hairpins con mayores tallos y menores loops.

En el alineamiento con tRNAs sería probablemente útil separar la base de datos de acuerdo a los distintos aminoácidos, o de acuerdo a los distintos anticodones, de modo de ver si hay anillos de A_0 que se "especialicen", siendo más cercanos a algunos tipos de tRNA. Es posible que un refinamiento de este tipo deba esperar a que se disponga de un volumen mayor de información, pues al descomponer la base de datos actualmente disponible los conjuntos (por anticodón, por ejemplo) resultan demasiado pequeños como para ofrecer una buena representatividad.

Tal vez sea interesante además el hecho de que los mejores alineamientos se logren con tRNAs mitocondriales: ¿qué pueden tener estos tRNAs de especial? Tal vez se podría argumentar que las mitocondrias, descendientes de bacterias muy antiguas, pueden haber conservado mejor las secuencias originales, pues al vivir inmersas en la célula eucariota, y protegidas por ésta, no habrían estado expuestas a las presiones que actuaron sobre sus primas que vivieron en libertad; o al menos, suponiendo que sí hubo presión, lo que no hubo fue un número tan grande de generaciones. Se trata sólo de una conjetura, pero susceptible de ser abordada mediante estudios filogenéticos. (Un posible argumento en contra es que precisamente en mitocondrias es donde se dan los cambios más bruscos en los tRNAs: omisión de todo el loop TΨC, o alteraciones del código genético; sin embargo, estos cambios violentos podrían estar concentrados en unos pocos linajes y ser atribuíbles a circunstancias puntuales –nuevamente, se trata de conjeturas).

Un tema que no se pudo abordar, por la dificultad para conseguir suficiente información ordenada, fue el de los intrones del tRNA. Estos intrones son interesantes, pues pertenecen al grupo I, con actividad riboenzimática que cataliza su propia edición, y codifican pequeños RNAs necesarios para la formación del ribosoma. Son por lo tanto candidatos probables para provenir del origen mismo de la maquinaria celular, y el hecho de que dentro de su secuencia esté contenido *el codón* del tRNA en que están insertos, sugiere que ya estaban presentes en los tiempos en que la forma misma del tRNA estaba naciendo (y por lo tanto, se acercan a la etapa en que los anillos habrían dado paso a la maquinaria de traducción).

Existen finalmente algunas tareas pendientes y de sumo interés que escapan definitivamente al ámbito de esta Tesis, por tratarse de trabajo de laboratorio. En primer lugar, es fundamental volver a visitar todas estas ideas a la luz de información real respecto a las afinidades entre codones (o anticodones, o pares codón/anticodón) con los aminoácidos que codifican. Este trabajo ganaría mucho si su punto de partida fuese una tabla con datos

experimentales respecto a esas afinidades, y no el código genético (el cual se usó como si fuese esa tabla, en virtud de un apego –provisorio– a una versión fuerte de la teoría estereoquímica).

Por otro lado, sería de interés estudiar las propiedades *in vitro* de la secuencia AL. ¿Quién sabe si acaso no es una ribozima natural, o un sustrato propicio para aminoacilación? El reciente desarrollo de técnicas para la generación y estudio de cadenas (*catenanes*) y anillos (*rotaxanes*) de RNA debería posibilitar experimentos en este sentido.

En resumen, el futuro de las premisas y consideraciones que dan forma a nuestro estudio y argumentación estará determinado por un exhaustivo trabajo computacional, análisis más detallados de los datos existentes, la recopilación de más datos a través de los proyectos de secuenciamiento en curso y futuros, y sobre todo, por información nueva sobre las propiedades químicas de las moléculas y sus interacciones. Si bien la argumentación en su forma actual resulta bastante elocuente, exige ser revisada en la medida en que estos nuevos elementos estén disponibles.

En lo inmediato, la tarea principal será ensamblar este nueva pieza del puzzle con las otras ya existentes respecto al origen de la maquinaria celular, muchas de las cuales deberían hallarse entre las expuestas en el Capítulo 3. En otras palabras, aún hay que escribir la historia (aquí apenas esbozada) de cómo los anillos permitieron el paso del mundo prebiótico al mundo vivo.

Capítulo 9

Glosario

Incluyo aquí un brevísimo glosario de algunos términos usados en esta parte de la Tesis. No pretende en ningún caso dar definiciones rigurosas; más bien es un ayuda-memoria para lectores ajenos al área, que pese a haber leído el Capítulo 2, se encuentren más tarde con algún término cuyo significado no recuerdan. También puede serle útil a algún biólogo perplejo que quiera ver qué fue lo que yo malentendí de algún concepto.

aminoacilación: Proceso que liga un aminoácido a un tRNA.

autocatálisis: Propiedad de algunos sistemas, en que los productos de unas reacciones catalizan la realización de otras, de manera cerrada, de modo que el sistema se "autoproduce".

autopoiesis: Noción de vida minimal definida por Varela y Maturana. A la propiedad de la autocatálisis se le agrega el requisito de la existencia de algún tipo de frontera o membrana, producida también por el sistema.

base: (O "base nitrogenada"). Son las moléculas, de cinco tipos distintos, que definen el alfabeto de los ácidos nucleicos. Con frecuencia hemos abusado del término y lo hemos usado para referirnos a los nucleótidos de los cuales las bases forman parte.

catálisis: Facilitamiento (aceleración o provocación) de una reacción química.

codón: Tripleta de bases en el DNA, que codifica un aminoácido o una señal de término.

cuasiespecie: Conjunto de secuencias parecidas, todas ellas variantes de una cierta "secuencia maestra", representante de la cuasiespecie. Noción introducida por Eigen.

enzima: Molécula (por lo general proteína) con actividad catalítica.

fósil molecular: Parte conservada en alguna estructura molecular, que por ser compartida por muchas especies, se considera un residuo de épocas pasadas.

hairpin: Forma adoptada por un RNA, plegándose sobre sí mismo, formando un "tallo" (dos cadenas apareadas) unido a un loop (un puente entre dos extremos cercanos de ambas cadenas).

hibridación: Asociación de una cadena de ácido nucleico con otra, a través de los pareos de Watson-Crick.

hidrofóbico, hidrofílico: Propiedad de los aminoácidos, que determina en gran medida la forma que adoptan las proteínas.

hiperciclo: Conjunto autocatalítico de cuasiespecies. Noción introducida por Eigen.

intrón: Segmento inserto dentro de un gen, que no se traduce, sino que se corta durante el proceso de *splicing*.

mitocondria: Organelo ubicado en las células eucariotas; es casi una "sub-célula".

mRNA: RNA mensajero; es la copia que se hace de algún trozo del DNA, para luego ser editado y traducido en una proteína.

nucleótido: Molécula formada por una base, un azúcar y un fosfato. Son los monómeros de los ácidos nucleicos. En este texto se ha hablado por lo general de "base" para referirse a lo que en realidad son nucleótidos, para simplificar.

pirimidina: Tipo de base usado en los ácidos nucleicos; incluye a la citosina (**c**), a la timina (**t**) y al uracilo (**u**), y lo hemos denotado con la letra Y.

polímero: Macromolécula formada por una cadena de unidades menores (monómeros). Ejemplos: proteínas, ácidos nucleicos, lípidos, polisacáridos.

purina: Tipo de base usado en los ácidos nucleicos; incluye a la guanina (**g**) y a la adenina (**a**), y lo hemos denotado con la letra R.

ribosoma: Estructura, formada por RNA, en la cual se lleva a cabo la traducción.

ribozima: Enzima formada por RNA.

stop: Codón de término (**uga**, **uaa** o **uag**).

traducción: Proceso a través del cual se lee un mRNA y se construye la proteína que éste codifica.

transcripción: Copia de segmentos de DNA en mRNA, para ser luego decodificado.

transferencia horizontal: Paso de un segmento de ácido nucleico desde una especie a otra, a través de la filtración de las paredes celulares.

transición: Mutación que reemplaza una base por otra del mismo tipo (purina por purina, o pirimidina por pirimidina). Es más frecuente, en la replicación, que la transversión.

transversión: Mutación que reemplaza una base por una de otro tipo (purina por pirimidina, o viceversa).

tRNA: RNA de transferencia; es un RNA que adopta forma de trébol, dejando en un extremo un anticodón y en el otro una cadena a la que se liga un aminoácido (correspondiente al anticodón). De esta forma, juega el rol principal en la traducción.

Capítulo 10

Bibliografía

Aita, T., Urata, S., y Husimi, Y. (2000). From amino acid landscape to protein landscape: Analysis of genetic codes in terms of fitness landscape. *J. Mol. Evol.*, 50:313–323.

Alberti, S. (1997). The origin of the genetic code and protein synthesis. *J. Mol. Evol.*, 45:352–358.

Aminorvin, R. (1997). An analysis of the metabolic theory of the origin of the genetic code. *J. Mol. Evol.*, 44:473–476.

Andersson, S. y Kurland, C. (1991). An extreme codon preference strategy: Codon reassignment. *Mol. Biol. Evol.*, 8:530–544.

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

Ardell, D. y Sella, G. (2001). On the evolution of redundancy in genetic codes. *J. Mol. Evol.*, 53:269–281.

Arquès, D., Lacan, J., y Michel, C. (2002). Identification of protein coding genes in genomes with statistical functions based on the circular code. *Biosystems*, 66:73–92.

Arquès, D. y Michel, C. (1995). A possible code in the genetic code. En Mayr, E. y Puech, C., editores, *STACS 95: Symposium on Theoretical Aspects of Computer Science*, volumen 900 de *Lect. Notes Comput. Sc.*, páginas 640–651. Springer Verlag, Berlin.

Arquès, D. y Michel, C. (1996). A complementary circular code in the protein coding genes. *J. Theor. Biol.*, 182:45–58.

Assouline, S., Nir, S., y Lahav, N. (2001). Simulation of non-enzymatic template-directed synthesis of oligonucleotides and peptides. *J. Theor. Biol.*, 208:117–125.

Atkins, P. y de Paula, J. (2002). *Physical Chemistry*, capítulo 22. Macromolecules and aggregates. Freeman & Co, New York. 7. Edición.

Bass, B. (2001). *RNA Editing*. Oxford University Press, New York.

Benson, D. et al. (2002). Genbank. *Nucleic Acids Res.*, 30:17–20.

Binder, S., Schuster, W., Grienenberger, J., Weil, J., y Brennicke, A. (1990). Genes for Gly-, His-, Lys-, Phe-, Ser- and Tyr-tRNA are encoded in œnothera mitochondrial DNA. *Curr. Genet.*, 17:353–358.

Blalock, J. (1990). Complementarity of peptides specified by "sense" and "antisense" strands of DNA. *Trends Biotechnol.*, 8:140–144.

Blalock, J. y Bost, K. (1986). The binding of peptides that are specified by complementary RNAs. *Biochem. J.*, 234:679–683.

Blalock, J. y Smith, E. (1984). Hydropathic anti-complementarity of amino acids based on the genetic code. *BBRC*, 121:203–207.

Böhler, C., Nielsen, P., y Orgel, L. (1995). Template switching between PNA and RNA oligonucleotides. *Nature*, 376:578–581.

Brooks, D., Fresco, J., Lesk, A., y Singh, M. (2002). Evolution of amino acid frequencies in proteins over deep time: Inferred order of introduction of amino acids into the genetic code. *Mol. Biol. Evol.*, 19:1645–1655.

Bujdák, J. y Rode, B. (1996a). The effect of smectite composition on the catalysis of peptide bond formation. *J. Mol. Evol.*, 43:326–333.

Bujdák, J. y Rode, B. (1996b). Silica, alumnia, and clay-catalyzed alanine peptide bond formation. *J. Mol. Evol.*, 45:457–466.

Cairns-Smith, A. (1982). *Genetic Takeover and the Mineral Origins of Life*. Cambridge University Press, Cambridge.

Chechetkin, V. (2003). Block structure and stability of the genetic code. *J. Theor. Biol.*, 222:177–188.

Chipens, G. (1991a). The evolution of the structures of amino acid families. *Zh. Evol. Biokhim. Fiziol.*, 27:513–521.

Chipens, G. (1991b). Hidden symmetry of the genetic code and laws of amino acid interaction. *Bioorg. Khim.*, 17:1335–1346.

Clarke, B. y Blalock, J. (1991). Characteristics of peptides specified by antisense nucleic acids. En vander Krol, A. y Mol, J., editores, *Applications of Antisense Nucleic Acids*, páginas 169–185. Marcel Dekker, Inc., New York.

Crick, F. (1968). The origin of the genetic code. *J. Mol. Biol.*, 38:367–379.

Crick, F., Brenner, S., Klug, A., y Pieczenik, G. (1976). A speculation on the origin of protein synthesis. *Origins of Life*, 7:389–.

Cullman, G. y Labouygues, J. (1982). Le code Baudot, trame du code génetique. *C.R. Acad. Sci. III-Vie*, 296:767–770.

Demongeot, J. (1975). *Au sujet de quelques modèles stochastiques appliqués à la biologie*. Ph.d. Thèse, Université Scientifique et Médicale de Grenoble.

Demongeot, J. (1978). Sur la possibilité de considérer le code génétique comme un code à enchaînement dégéneré. *Revue de Biomaths*, 62:61–66.

Demongeot, J. y Besson, J. (1983). Code génétique et codes à enchaînement. *C.R. Acad. Sci. III-Vie*, 296:807–810.

Demongeot, J. y Besson, J. (1996). The genetic code and cyclic codes. *C.R. Acad. Sci. III-Vie*, 319:443–451.

Di Giulio, M. (1995). Was it an ancient gene codifying for a hairpin RNA that, by means of direct duplication, gave rise to the primitive tRNA molecule? *J. Theor. Biol.*, 177:95–101.

Di Giulio, M. (1999). The non-monophyletic origin of the tRNA molecule. *J. Theor. Biol.*, 197:403–414.

Di Giulio, M. y Medugno, M. (2000). The robust statistical bases of the coevolution theory of genetic code origin. *J. Mol. Evol.*, 50:258–263.

Diaz-Lazcoz, Y. (1998). Evolution of genes, evolution of species: The case of aminoacyl-tRNA synthetases. *Mol. Biol. Evol.*, 15:1548–1561.

Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science*, 284:2124–2128.

Doudna, J. y Cech, T. (2002). The chemical repertoire of natural ribozymes. *Nature*, 418:222–228.

Dufton, M. (1997). Genetic code synonym quotas and amino acid complexity: Cutting the cost of proteins? *J. Theor. Biol.*, 187:165–173.

Dunnill, P. (1966). Triplet nucleotideamino acid pairing: a stereochemical basis for the division between protein and nonprotein amino acids. *Nature*, 210:1267–1268.

Dyson, F. (1985). *Origins of Life*. Cambridge University Press, Cambridge.

Eigen, M. (1992). *Step Towards Life: A Perspective on Evolution*. Oxford University Press, New York.

Eigen, M., Gardiner, W., Schuster, P., y Winkler-Oswatitsch, R. (1981). The origin of genetic information. *Sci. Am.*, 244:88–118.

Eigen, M. y Schuster, P. (1978). The hypercycle. a principle of natural self-organization. Part C: The realistic hypercycle. *Naturwissenschaften*, 65:314–369.

Felden, B., Florentz, C., Westhof, E., y Giegé, R. (1998). Transfer RNA identity rules and conformation of the tyrosine tRNA-like domain of BMV RNA imply additional charching by histidine and valine. *Biochem. Bioph. Res. Co.*, 243:426–434.

Fitch, W. y Upper, K. (1988). The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. *Cold Spring Harb. Sym.*, 52:759–767.

Forsdyke, D. (1995). Sense in antisense? *J. Mol. Evol.*, 41:582–586.

Francklyn, C. y Schimmel, P. (1990). Enzymatic aminoacylation of an eight-base-pair microhelix with histidine. *P. Natl. Acad. Sci. USA*, 87:8655–8659.

Freeland, S. (2001). The Darwinian genetic code: an adaptation for adapting? Preprint.

Freeland, S., Knight, R., Landweber, L., y Hurst, L. (2000). Early fixation of an optimal genetic code. *Mol. Biol. Evol.*, 17:511–518.

Gao, K. y Orgel, L. (2000). Polyphosphorylation and non-enzymatic template-directed ligation of oligonucleotides. *Origins Life Evol. B.*, 30:45–51.

Garcia, I. y Weeks, K. (2003). Small structural costs for evolution from RNA to RNP-based catalysis. *J. Mol. Biol.*, 331:57–73.

Gatlin, L. (1968). The information content of DNA ii. *J. Theor. Biol.*, 18:181–194.

Gilbert, W. (1986). The RNA world. *Nature*, 319:618.

Gilis, D., Massar, S., Cerf, N., y Rooman, M. (2001). Optimality of the genetic code with respect to protein stability and amino-acid frecuencies. *Genome Biol.*, 2(0049).

Gogarten, J., Doolittle, W., y Lawrence, J. (2002). Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.*, 19:2226–2238.

Grafstein, D. (1983). Stereochemical origins of the genetic code. *J. Theor. Biol.*, 105:157–174.

Haig, D. y Hurst, L. (1991). A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.*, 33:412–417.

Harris, L., Sullivan, M., y Hatfield, D. (1999). Directed molecular evolution. *Origins Life Evol. B.*, 29:425–435.

Harshey, R. (2003). Lecture on RNA structure.
`http://www.esb.utexas.edu/rasika/695D/rna_structure/rna_html/RNA%20Structure.htm`.

Hartman, H. (1984). Speculation on the evolution of the genetic code iii: the evolution of tRNA. *Origins of Life*, 14:643–648.

Hartman, H. (1995). Speculation on the evolution of the genetic code iv: The evolution of the aminoacyl-tRNA synthetases. *Origins Life Evol. B.*, 25:265–269.

Hendry, L. et al. (1981). First approximation of a stereochemical rationale for the genetic code based on topography and physiochemical properties of "cavities" constructed from models of DNA. *P. Natl. Acad. Sci. USA*, 78:7440–7444.

Hendry, L. et al. (1995). A stereochemical rationale for the genetic code derived from complementary fit of amino acids into cavities formed in codon/anticodon sequences in double stranded DNA: Further evidence based upon noncomplementarity of untranslated amino acids. *The World Wide Web Journal of Biology*, 1(3). `http://www.epress.com/w3jbio/vol1/hendry/hendry.html`.

Hipps, D. et al. (1995). Operational RNA code for amino acids: Species-specific aminoacylation of minihelices switched by a single nucleotide. *P. Natl. Acad. Sci. USA*, 92:5550–5552.

Hobish, M., Wickramasinghe, N., y Ponnamperuma, C. (1995). Direct interaction between amino acids and nucleotides as a possible physicochemical basis for the origin of the genetic code. *Adv. Space Res.*, 15:365–382.

Houen, G. (1999). Evolution of the genetic code: the nonsense, antisense, and antinonsense codes make no sense. *Biosystems*, 54:39–46.

Ikawa, Y., Tsuda, K., Matsumura, S., Atsumi, S., y Inoue, T. (2003). Putative intermediary stages for the molecular evolution from a ribozyme to a catalytic RNP. *Nucleic Acid Res.*, 31:1488–1496.

Ikehara, K. (2002a). A novel theory on the origin of the genetic code: A GNC-SNS hypothesis. *J. Mol. Evol.*, 54:530–538.

Ikehara, K. (2002b). Origins of gene, genetic code, protein and life: comprehensive view of life systems from a GNC-SNS primitiv genetic code hypothesis. *J. Bioscience*, 27:165–186.

Jackson, S., Cannone, J., Lee, J., Gutell, R., y Woodson, S. (2002). Distribution of rRNA introns in the three-dimensional structure of the ribosome. *J. Mol. Biol.*, 323:35–52.

Jarpe, M. y Blalock, J. (1994). Complementary peptides: applications of the molecular recognition theory to peptide and protein purification and design. En Basava, C. y Anantharamaiah, G., editores, *Peptides: Design, Synthesis, and Biological Activity*, páginas 165–179, Boston. Birkhauser.

Jäschke, A. (2001). Artificial ribozymes and deoxyribozymes. *Curr. Opin. Struc. Biol.*, 11:321–326.

Jeffares, D., Poole, A., y Penny, D. (1998). Relics from the RNA world. *J. Mol. Evol.*, 46:18–36.

Johnston, W., Unrau, P., Lawrence, M., Glasner, M., y Bartel, D. (2001). RNA-catalyzed RNA polymerization: Accurate and general RNA-templated primer extension. *Science*, 292:1319–1325.

Jolivet, R. y Rothen, F. (2001). Peculiar symmetry of DNA sequences and evidence suggesting its evolutionary origin in a primeval genetic code. En *First European Workshop on Exo-/Astro-Biology*, volumen 496 de *ESA SP*, páginas 173–176, Frascati.

Joyce, G. (2002). The antiquity of RNA-based evolution. *Nature*, 418:214–221.

Jukes, T. (1996). On the prevalence of certain codons in genes for proteins. *J. Mol. Evol.*, 42:377–381.

Kargupta, H. (2000). A striking property of genetic code-like transformations. *Complex Systems*, 13:1–32.

Karlin, S. y Mrázek, J. (1997). Compositional differences within and between eukaryotic genomes. *P. Natl. Acad. Sci. USA*, 94:10227–10232.

Kauffman, S. (1993). *Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, New York.

Kauffman, S. (1996). Self-replication: Even peptides do it. *Nature*, 382.

Knight, R., Freeland, S., y Landweber, L. (1999). Selection, history and chemistry: the three faces of the genetic code. *Trends Biochem. Sci.*, 24:241–247.

Knight, R., Freeland, S., y Landweber, L. (2001a). Rewiring the keyboard: evolvability of the genetic code. *Nat. Rev. Genet.*, 2:49–58.

Knight, R. y Landweber, L. (1998). Rhyme or reason: RNA-arginine interactions and the genetic code. *Chem. Biol.*, 5:R215–R220.

Knight, R. y Landweber, L. (2000). The early evolution of the genetic code. *Cell*, 101:569–572.

Knight, R., Landweber, L., y Yarus, M. (2001b). How mitochondria redefine the code. *J. Mol. Evol.*, 53:299–313.

Kochavi, E., Bar-Nun, A., y Fleminger, G. (1997). Substrate-directed formation fo small biocatalysts under prebiotic conditions. *J. Mol. Evol.*, 45:342–351.

Konecny, J., Schöniger, M., y Hofacker, L. (1995). Complementary coding conforms to the primeval coma-less code. *J. Theor. Biol.*, 173:263–270.

Krakauer, D. y Jansen, V. (2002). Red queen dynamics of protein translation. *J. Theor. Biol.*, 218:97–109.

Labouygues, J. y Figureau, A. (1984). The logic of the genetic code: synonyms and optimality against effects of mutations. *Origins of Life*, 14:685–692.

Landweber, L. (1999). Experimental RNA evolution. *Trends Ecol. Evol.*, 14:353–358.

Larkin, D. et al. (2002). Identification of essential domains for *eschericia coli* tRNAleu aminoacylation and amino acid editing using minimalist RNA molecules. *Nucleic Acid Res.*, 30:2103–2113.

Lazcano, A. (1997). Chemical evolution and the primitive soup: Did Oparin get it all right? *J. Theor. Biol.*, 184:219–223.

Le Touze, G. (1995). *Analyse de l'organization du génome de Saccharomyces cerevisiae et recherche d'un code cyclique non-dégénéré ancestral à l'origine du code génétique actuel*. Thesis, dea de biomathématiques, Universités Paris VI - Paris VII.

Lee, D., Granka, J., Martínez, J., Severin, K., y Ghadiri, M. (1996). A self-replicating peptide. *Nature*, 382:525–528.

Lee, D., Severin, K., y Ghadiri, M. (1997a). Autocatalytic networks: the transition from molecular self-replication to ecosystems. *Curr. Opin. Chem. Biol.*, 1:491–496.

Lee, D., Severin, K., Yokobayashi, Y., y Ghadiri, M. (1997b). Emergence of symbiosis in peptide self-replication through a hypercyclic network. *Nature*, 390:591–594.

Lee, N., Bessho, Y., Wei, K., Szostak, J., y Suga, H. (2000). Ribozyme-catalyzed tRNA aminoacylation. *N. Struct. Biol.*, 7:28–33.

Leroy, L. (1993). *L'Origin de la vie*. Biocosmos Editions, Larcay.

Lewin, B. (2002). *Genes VII*. Oxford University Press, New York.

Luisi, P. (1993). Defining the transition to life: Self-replicating bounded structures and chemical autopoiesis. En Stein, W. y Varela, F., editores, *Thinking About Biology*, SFI Studies in the Sciences of Complexity. Addison-Wesley.

Luo, L. y Li, X. (2002a). Coding rules for amino acids in the genetic code: the genetic code is a minimal code of mutational deterioration. *Origins Life Evol. B.*, 32:23–33.

Luo, L. y Li, X. (2002b). Construction of genetic code from evolutionary stability. *Biosystems*, 65:83–97.

Maher, B. (2002). Uprooting the tree of life. *The Scientist*, 16:18/26.

Maizels, N. y Weiner, A. (1995). Phylogeny from function: The origin of tRNA is in replication, not translation. En Fitch, W. y Ayala, F., editores, *Tempo and Mode in Evolution: Genetics and Paleontology 50 Years After Simpson*. National Academy of Science.

Maizels, N., Weiner, A., Yue, D., y Shi, P. (1999). New evidence for the genomic tag hypothesis: Archaeal CCA-adding enzymes and tDNA substrates. *Biol. Bull.*, 196:331–333.

Majerfeld, I. y Yarus, M. (1998). Isoleucine: RNA sites with essential coding sequences. *RNA*, 4:471–478.

Mathews (1988). No code for recognition. *Nature*, 335:294–295.

Matzura, O. y Wennborg, A. (1996). RNAdraw: an integrated progrgam for RNA secondary structure calculation and analysis under 32-bit Microsoft Windows. *Bioinformatics*, 12:247–249.

Maynard Smith, J. y Szathmáry, E. (1995). *The Major Transitions in Evolution*. Freeman, Oxford.

Maynard Smith, J. y Szathmáry, E. (2001). *Ocho hitos de la evolución*. Tusquets Editores S.A., Barcelona.

McGinness, K. y Joyce, G. (2002). Continuous in vitro evolution of a ribozyme that catalyzes three successive nucleotidyl addition reactions. *Chem. Biol.*, 9:585–596.

McGinness, K. y Joyce, G. (2003). In search of an RNA replicase ribozyme. *Chem. Biol.*, 10:5–14.

Mekler, L. (1969). On the specific selective interaction between amino acid residues of polypeptide chains. *Biofizika*, 14:581–584.

Mellersh, A. (1993). A model for the prebiotic synthesis of peptides which throws light on the origin of the genetic code and the observed chirality of life. *Origins Life Evol. B.*, 23:261–274.

Mellersh, A. y Wilkinson, A.-S. (2000). RNA bound to a solid phase can select an amino acid and facilitate subsequent amide bond formation. *Origins Life Evol. B.*, 30:3–7.

Miramontes, P. et al. (1995). Structural and thermodynamic properties of DNA uncover different evolutionary histories. *J. Mol. Evol.*, 40:698–704.

Moreira, A. (1999). Tratamiento matemático de la traducción inversa de proteínas. Memoria de titulación, Ing. Civil Matemática, Depto. de Ing. Matemática, U. de Chile.

Moreira, A. (2003). Genetic algorithms for the imitation of genomic styles in protein backtranslation. To appear in Theor. Comput. Sci.

Moreira, A. y Maass, A. (2003). TIP: Protein backtranslation aided by genetic algorithms. *Bioinformatics*. To appear.

Nakamura, Y., Gojobori, T., y Ikemura, T. (2000). Codon usage tabulated from the international DNA sequence databases; its status for the year 2000. *Nucleic Acids Res.*, 28:292.

Nashimoto, M. (2001). The RNA/protein symmetry hypothesis: Experimental support for reverse translation of primitive proteins. *J. Theor. Biol.*, 209:181–187.

Nielsen, P. (1993). Peptide nucleic acid (PNA): a model structure of the primordial genetic material? *Origins Life Evol. B.*, 23:323–327.

Nisbet, E. y Sleep, N. (2001). The habitat and nature of early life. *Nature*, 409:1083–1091.

Ohnishi, K. (2000). Origin of mRNAs and genetic code by means of hierarchical sociogenesis of tRNA-riboorganisms. En *Fourth International Conference on Emergence*, Odense, Denmark.

Osawa, S. et al. (1992). Recent evidence for evolution of the genetic code. *Microbiol. Rev.*, 56:229–264.

Osawa, S. y Jukes, T. (1995). On codon reassignment. *J. Mol. Evol.*, 41:247–249.

Pasqual, N., Gallagher, M., Aude-Garcia, C., Loidice, M., Thuderoz, F., Demongeot, J., Ceredig, R., Marche, P., y Jouvin-Marche, E. (2002). Quantitative and qualitative changes in ADV-AJ rearrangements during mouse thymocytes differentiation: implications for a limited TCR ALPHA chain repertoire. *J. Exp. Med.*, 196:1163–1174.

Paul, N. y Joyce, G. (2002). A self-replicating ligase ribozyme. *P. Natl. Acad. Sci. USA*, 99:12733–12740.

Pelc, S. y Welton, M. (1966). Stereochemical relationship between coding triplets and amino acids. *Nature*, 209:868–872.

Penny, D., Hendy, M., y Poole, A. (2003). Testing fundamental evolutionary hypotheses. *J. Theor. Biol.*, 223:377–385.

Pirillo, G. (2001). Remarks on the Arquès-Michel code. *Riv. Biol.-Biol. Forum*, 94:327–330.

Poole, A., Jeffares, D., y Penny, D. (1999). Early evolution: prokaryotes, the new kids on the block. *BioEssays*, 21:880–889.

Porschke, K. (1985). Differential effect of amino acid residues on the stability of double helices formed from polyribonucleotides and its possible relation to the evolution of the genetic code. *J. Mol. Evol.*, 21:192198.

Rodin, S., Ohno, S., y Rodin, A. (1993). Transfer RNAs with complementary anticodons: could they reflect early evolution of discriminative genetic code adaptors? *P. Natl. Acad. Sci. USA*, 90:4723–4727.

Rodin, S., Rodin, A., y Ohno, S. (1996). The presence of codon-anticodon pairs in the acceptor stem of tRNAs. *P. Natl. Acad. Sci. USA*, 93:4537–4542.

Ronneberg, T., Landweber, L., y Freeland, S. (2000). Testing a biosynthetic theory of the genetic code: Fact or artifact? *P. Natl. Acad. Sci. USA*, 97:13690–13695.

Root-Bernstein, R. y Holsworth, D. (1998). Antisense peptides: A critical mini-review. *J. Theor. Biol.*, 190:107–119.

Saito, H., Kourouklis, D., y Suga, H. (2001). An *in vitro* evolved precursor tRNA with aminoacylation activiy. *EMBO J.*, 20:1797–1806.

SaksLab (2003). `http://biology.uoregon.edu/Biology_WWW/People/Saks/lab/`.

Salehi-Ashtiani, K. y Szostak, J. (2001). In vitro evolution suggests multiple origins for the hammerhead ribozyme. *Nature*, 414:82–84.

Schimmel, P. (1996). Origin of the genetic code: A needle in the haystack of tRNA sequences. *P. Natl. Acad. Sci. USA*, 1996:4521–4522.

Schimmel, P. y Henderson, B. (1994). Possible role of aminoacyl-RNA complexes in noncoded peptide synthesis and origin of coded synthesis. *P. Natl. Acad. Sci. USA*, 91:11283–11286.

Schultz, D. y Yarus, M. (1994). Transfer RNA mutation and the malleability of the genetic code. *J. Mol. Biol.*, 235:1377–1380.

Schwartz, A. (1997). Speculation on the RNA precursor problem. *J. Theor. Biol.*, 187:523–527.

Segré, D., Ben-Eli, D., y Lancet, D. (2000). Compositional genomes: Prebiotic information transfer in mutually catalytic noncovalent assemblies. *P. Natl. Acad. Sci. USA*, 87:4112–4117.

Segré, D. et al. (2001). The lipid world. *Origins Life Evol. B.*, 31:119–145.

Seligmann, H. y Amzallag, G. (2002). Chemical interactions between amino acid and RNA: multiplicity of the levels of specificity explains origin of the genetic code. *Naturwissenschaften*, 89:542–551.

Sella, G. y Ardell, D. (2002). The impact of message mutation on the fitness of a genetic code. *J. Mol. Evol.*, 54:638–651.

Severin, K., Lee, D., Kennan, A., y Ghadiri, M. (1997a). A synthetic peptide ligase. *Nature*, 386:706–709.

Severin, K., Lee, D., Martinez, J., y Ghadiri, M. (1997b). Peptide self-replication via template-directed ligation. *Chem. Eur. J.*, 3:1017–1024.

Singer, M. y Berg, P. (1992). *Genes & Genomes*. Editions Vigot, Paris.

Soto, M. A. y Tohá, J. (1985). A hardware interpretation of the evolution of the genetic code. *Biosystems*, 18:209–215.

Sowerby, S. y Heckl, W. (1998). The role of self-assembled monolayers of the purine and pyrimidine bases in the emergence of life. *Origins Life Evol. B.*, 28:283–310.

Sowerby, S., Petersen, G., y Holm, N. (2002). Primordial coding of amino acids by adsorbed purine bases. *Origins Life Evol. B.*, 32:35–46.

Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A., y Steinberg, S. (1998). Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, 26:148–153.

Stevenson, D. (2002). Co-evolution of the genetic code and ribozyme replication. *J. Theor. Biol.*, 217:235–253.

Strazewski, P. (1994). *Recent Studies on the Synthesis, Structure and Function of Nucleic Acids*. Habilitationsschrift, Universität Basel.

Syvanen, M. (2002). Recent emergence of the modern genetic code. *Trends Genet.*, 18:245–248.

Szathmáry, E. (1993). Coding coenzyme handles: A hypothesis for the origin of the genetic code. *P. Natl. Acad. Sci. USA*, 90:9916–9920.

Szathmáry, E. (1999). The origin of the genetic code: amino acid as cofactors in an RNA world. *Trends Genet.*, 15:223–229.

Szathmáry, E. y Demeter, L. (1987). Group selection of replicators and the origin of life. *J. Theor. Biol.*, 128:463–486.

Szostak (2003). `http://xanadu.mgh.harvard.edu/szostakweb/web.2`.

Tanaka, T. y Kikuchi, Y. (2001). Origin of the cloverleaf shape of transfer RNA - the double-hairpin model: Implication for the role of tRNA intron and the long extra loop. *Viva Origino*, 29:134–142.

Tohá, J., Donoso, R., Estay, M., y Díaz-Valdés, J. (1989). Codon evolution and conservation of the reading phase in genetic code translation. *Med. Hypotheses*, 30:265–269.

Tracqui, P. y Demongeot, J., editores (2003). *Élements de Biologie à l'usage d'autres disciplines*, Grenoble. Grenoble Sciences.

Trifonov, E. (2000). Consensus temporal order of amino acids and evolution of a triplet code. *Gene*, 261:139–151.

Tropsha, A., Kizer, J., y Chaiken, I. (1992). Making sense from antisense: a review of experimental data and developing ideas on sense-antisense peptide recognition. *J. Mol. Recognit.*, 5:43–54.

Varela, F. (2000). *El fenómeno de la vida*. Dolmen Ediciones, Santiago.

Varela, F., Maturana, H., y Uribe, R. (1974). Autopoiesis: The organization of the living systems, its characterization and a model. *Biosystems*, 5:187–196.

Wächtershäuser, G. (1994). Life in a ligand sphere. *P. Natl. Acad. Sci. USA*, 89:4283–4287.

Wächtershäuser, G. (2000). Life as we don't know it. *Science*, 289:1307–1308.

Weber, A. y Lacey, J. (1978). Genetic code correlations: Amino acids and their anticodonic nucleotids. *J. Mol. Evol.*, 11:195–210.

Weil, G., Heus, K., Faraut, T., y Demongeot, J. (1995). The genetic code as a constraint satisfaction problem. Preprint.

Weiner, A. y Maizels, N. (1987). 3′ terminal tRNA-like structures tag genomic RNA molecules for replication: Implications for the origin of protein synthesis. *P. Natl. Acad. Sci. USA*, 84:7383–7387.

Weiner, A. y Maizels, N. (1999). The genomic tag hypothesis: Modern viruses as molecular fossils of ancient strategies for genomic replication, and clues regarding the origin of protein synthesis. *Biol. Bull.*, 196.

Wills, P. (1993). Self-organization of genetic coding. *J. Theor. Biol.*, 162:267–287.

Woese, C. (1965). Order in the genetic code. *P. Natl. Acad. Sci. USA*, 54:71–75.

Woese, C. (1967). *The genetic code: The Molecular Basis for Genetic Expression*. Harper & Row, New York.

Woese, C. (2002). On the evolution of cells. *P. Natl. Acad. Sci. USA*, 99:8742–8747.

Wong, J. (1980). Role of minimization of chemical distances between aminoacids in the evolution of the genetic code. *P. Natl. Acad. Sci. USA*, 77:1083–1086.

Wong, J. (1988). Evolution of the genetic code. *Microbiol. Sci.*, 5:174–181.

Wright, M. y Joyce, G. (1997). Continuous in vitro evolution of catalytic function. *Science*, 276:614–617.

Yarus, M. (2000). RNA-ligand chemistry: a testable source for the genetic code. *RNA*, 6:475–484.

Yokobori, S., Suzuki, T., y Watanabe, K. (2001). Genetic code variations in mitochondria: tRNA as a major determinant of genetic code plasticity. *J. Mol. Evol.*, 53:314–326.

Zhang, B. y Cech, T. (1997). Peptide bond formation by *in vitro* selected ribozymes. *Nature*, 390:96–100.

Zintzaras, E., Santos, M., y Szathmáry, E. (2002). "Living" under the challenge of information decay: The stochastic corrector model vs. hypercycles. *J. Theor. Biol.*, 217:167–181.

Capítulo 11

Anexo 1: Bibliografía adicional sobre el origen de la vida

Incluyo aquí algunas referencias adicionales para el tema del origen de la vida, aproximadamente organizadas según los tópicos tratados en la reseña del Capítulo 3.

C. Del Carpio et al, 1996. Sense-Antisense Homology Boxes in Proteins: Structural Motifs Encoded in the DNA? *Proceedings of the Seventh Workshop on Genome Informatics* 7:226-227.

J. Zull, S. Smith, 1990. Is genetic code redundancy related to retention of structural information in both DNA strands? *Trends Biochem. Sci.* 15:257–261.

W. Scott, 1998. RNA catalysis. *Curr. Opin. Struc. Biol.* 8:720–726.

S. Osborne, I. Matsumura, A. Ellington, 1997. Aptamers ad therapeutic and diagnostic reagents: problems and prospects. *Curr. Opin. Chem. Biol.* 1:5–9.

A.D. Ellington, 1993. Experimental Testing of Theories of an Early RNA World. *Method. Enzymol.* 224:646–664.

C. Frauendorf, A. Jäschke, 1998. Catalysis of Organic Reactions by RNA. *Angew. Chem. Int. Edit.* 37:1378–1381.

G.X. Zia et al, 2002. Directed evolution of novel polymerase activities: Mutation of a DNA polymerase into an efficient RNA polymerase. *P. Natl. Acad. Sci. USA* 99:6597-602.

J. Lehmann, 2000. Physico-chemical Constraints Connected with the Coding Properties of the Genetic System. *J. Theor. Biol.* 202, 129–144.

G. Fleminger, E. Kochavi, A. Bar-Nun, 1997. Substrate-Directed Formation of Catalytic Metallo-Oligopeptides. *At the Electronic Conference on Organometallic Chemistry, ECTOC-3.* `http://www.ch.ic.ac.uk/ectoc/ectoc-3/pub/044/ectoc.htm`.

S. Alberti, 1997. The Origin of the Genetic Code and Protein Synthesis. *J. Mol. Evol.* 45:352–358.

A. Gultyaev, F. van Batenburg, C. Pleij, 2002. Selective Pressures on RNA Hairpins in Vivo and In Vitro. *J. Mol. Evol.* 54:1–8.

M. Illangasekare, M. Yarus, 2002. Phenylalanine-Binding RNAs and Genetic Code Evolution. *J. Mol. Evol.* 54:298–311.

M. Rakocević, 1998. Whole-number Relations Between Protein Amino Acids and Their Biosynthetic Precursors. *J. Theor. Biol.* 191:463–465.

M. Rakocević, A. Jokić, 1996. Four Stereochemical Types of Protein Amino Acids: Synchronic Determination with Chemical Characteristics, Atom and Nucleon Number. *J. Theor. Biol.* 183:345–349.

K. Kawamura, F. Okamoto, 2001. Cyclization and Dimerization of hexanucleotides containing guanine and cytosine with water-soluble carbodiimide. *Viva Origino* 29:162–167.

G. Weberndorfer, 2002. Computational Models of the Genetic Code Evolution Based on Empirical Potentials. *Ph.D. Dissertation, Universität Wien.*

Grupo de Landweber

Los siguientes dos artículos están juntos (se discuten uno al otro):

S. Freeland, R. Knight, L. Landweber, 2000. Measuring adaptation within the genetic code. *Trends Biochem. Sci.* 25:44–45.

M. Di Giulio, 2000. The origin of the genetic code. *Trends Biochem. Sci.* 25:44.

T. Ronneberg, S. Freeland, L. Landweber, 2001. *Genview* and *Gencode*: a pair of programs to test theories of genetic code evolution. *Bioinformatics* 17:280–281.

S. Freeland, R. Knight, L. Landweber, 1999. Molecular Evolution: Do Proteins Predate DNA? *Science* 286: 690–692.

L. Landweber, P. Simon, T. Wagner, 1998. Ribozyme Engineering and Early Evolution. *Bioscience* 48:94–103.

L. Landweber, I. Pokrovskaya, 1999. Emergence of a dual-catalytic RNA with metal-specific cleavage and ligase activities: The spandrels of RNA evolution. *P. Natl. Acad. Sci. USA* 96:173–178.

R. Knight, L. Landweber, 1999. Is the genetic code really a frozen accident? New evidence from in vitro selection. *Ann. N. Y. Acad. Sci.* 870:408–10.

R. Knight, L. Landweber, 2000. Guilt by association: The arginine case revisited. *RNA* 6:499–510.

R. Knight, S. Freeland, L. Landweber, 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol.* 2:0010.

S. Freeland, L. Hurst, 1998. The Genetic Code is One in a Million. *J. Mol. Evol.* 47:238–248.

L. Landweber, W. Gilbert, 1994. Phylogenetic analysis of RNA editing: A primitive genetic phenomenon. *P. Natl. Acad. Sci. USA* 91:918–921.

A. Weiner, N. Maizels, 1997. The Genomic Tag Hypothesis: Modern Viruses as Molecular Fossils of Ancient Strategies for Genomic Replication, and Clues Regarding the Origin of Protein Synthesis.

N. Maizels, A. Weiner, D. Yue, P. Shi, 1997. New Evidence for the Genomic Tag Hypothesis: Archaeal CCA-Adding Enzymes and tDNA Substrates.

P. Shi, N. Maizels, A. Weiner, 1998. CCA addition by tRNA nucleotidyltransferase: polymerization without translocation? *EMBO J.* 17:3197–3206.

A. Kowald, T.B.L. Kirkwood, 1993. Accuracy of tRNA Charging and Codon:Anticodon Recognition; Relative Importance for Cellular Stability. *J. Theor. Biol.* 160:493–508.

P. Mucha et al, 2001. Anticodon domain methylated nucleosides of yeast $tRNA^{Phe}$ are significant recognition determinants in the binding of a phage display selected peptide. *4th Symposium on RNA Biology, Chapel Hill, North Carolina, 2001.*

M. Ibba et al, 2000. The Adaptor Hypothesis revisited. *Trends Biochem. Sci.* 25:311–316.

P. Schimmel, L.R. de Pouplana, 1999. Genetic code origins: Experiments confirm phylogenetic predictions and may explain a puzzle. *P. Natl. Acad. Sci. USA* 96:327–328.

P. Schimmel et al, 1993. [Review] An operational RNA code for amino acids and possible relationships to genetic code. *P. Natl. Acad. Sci. USA* 90:8763–8768.

L.R. de Pouplana et al, 1996. Evidence that two present-day components needed for the genetic code appeared after nucleated cells separated from bacteria. *P. Natl. Acad. Sci. USA* 93:166–170.

P. Schimmel, R. Alexander, 1998. Diverse RNA substrates for aminoacylation: Clues to origins? *P. Natl. Acad. Sci. USA* 95:10351–10353.

L.R. de Pouplana et al, 1998. Genetic code origins: tRNAs older than their synthetases? *P. Natl. Acad. Sci. USA* 95:11295–11300.

M. Szymanski et al, 2001. Aminoacyl-tRNA synthetases database. *Nucleic Acid Res.* 29:288–290.
The database is at
`http://biobases.ibch.poznan.pl/aars/`

A. Cavalcanti, R. Ferreira, 2001. On the relative content of G, C bases in codons of amino acids corresponding to Class I and II aminoacyl-tRNA synthetases. *Origins Life Evol. B.* 31:257–269.

E. Berthonneau, M. Mirande, 2000. A gene fusion event in the evolution of aminoacyl-tRNA synthetases. *FEBS Lett.* 470:300–304.

K. Yusim, S. Nir, N. Lahav, 2001. A Model for Proto-tRNA Loading. *J. Theor. Biol.* 208:109–116.

C. Myers et al, 2002. tRNA-like recognition of group I introns by a tyrosyl-tRNA synthetase. *P. Natl. Acad. Sci. USA* 99:2630–2635.

Más sobre el origen y evolución del tRNA

S. Rodin, S. Ohno, 1997. Four primordial models of tRNA-synthetase recognition, determined by the (G,C) operational code. *P. Natl. Acad. Sci. USA* 97:5183–5188.

F. Lustig et al, 1993. The nucleotide in position 32 of the tRNA anticodon loop determines ability of anticodon UCC to discriminate among glycine codons. *P. Natl. Acad. Sci. USA* 90:3343–3347.

M.J. Laforest, I. Roewer, B.F. Lang, 1997. Mitochondrial tRNAs in the lower fungus Spizellomyces punctatus: tRNA editing and UAG 'stop'codons recognized as leucine. *Nucleic Acid Res.* 25:626–632.

Y. Inagaki et al, 1998. Directionally Evolving Genetic Code: The UGA Codon from Stop to Tryptophan in Mitochondria. *J. Mol. Evol.* 47:378–384.

M. Lynch, 1996. Mutation accumulation in transfer RNAs: molecular evidence for Muller's reatchet in mitochondrial genomes. *Mol. Biol. Evol.* 13:209–220.

M. Lynch, 1997. Mutation Accumulation in Nuclear, Organelle, and Prokaryotic Transfer RNA Genes. *Mol. Biol. Evol.* 14:914–925.

P. León, 1998. Inhibition of Ribozymes by Deoxyribonucleotides and the Origin of DNA. *J. Mol. Evol.* 47:122–126.

Teorías adaptativas y biosintéticas

M.L. Chiusano, F. Alvarez-Valin, M. Di Giulio, G. DÓnofrio, G. Ammirato, G. Colonna, G. Bernardi, 2000. Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of the genetic code. *Gene* 261:65–69.

G. Cullman, J.M. Labouygues, 1983. Noise immunity of the genetic code. *Biosystems* 16:9–29.

M. Di Giulio, 1998. Reflections on the Origin of the Genetic Code: a Hypothesis. *J. Theor. Biol.* 191:191–196.

M. Di Giulio, 1997. On the Origin of the Genetic Code. *J. Theor. Biol.* 187:573–581.

M. Di Giulio et al, 1994. On the Optimization of the Physicochemical Distances between Amino Acids in the Evolution of the Genetic Code. *J. Theor. Biol.* 168:43–51.

M. Di Giulio, 1998. On the RNA World: Evidence in Favor of an Early Ribonucleopeptide World. *J. Mol. Evol.* 45:571–578.

M. Di Giulio, M. Medugno, 1998. The Historical Factor: The Biosynthetic Relationships Between Amino Acids and Their Physicochemical Properties in the Origin of the Genetic Code. *J. Mol. Evol.* 46:615–621.

M. Di Giulio, M. Medugno, 1999. Physicochemical Optimization in the Genetic Code Origin as the Number of Codified Amino Acids Increases. *J. Mol. Evol.* 49:1–10.

M. Di Giulio, 2000. Genetic Code Origin and the Strength of Natural Selection. *J. Theor. Biol.* 205:659–661.

O. Judson, 1999. The Genetic Code: What Is It Good For? An Analysis of the Effects of Selection Pressures on Genetic Codes. *J. Mol. Evol.* 49:539–550.

P.-F. Baisnée et al, 2001. Flexibility of the genetic code with respect to DNA structure. *Bioinformatics* 17:237–248.

J.-L. Jestin, A. Kempf, 1997. On the codon assignment of chain termination signals and the minimization of the effects of frameshift mutations.
arxiv:physics/9706019.

R. Amirnovin, 1997. An Analysis of the Metabolic Theory of the Origin of the Genetic Code. *J. Mol. Evol.* 44:473–476.

B. Davis, 1999. Evolution of the genetic code. *Prog. Biophys. Mol. Bio.* 72:157–243.

E. Trifonov, 1999. Glycine clock: Eubacteria first, Archae next, Protoctista, Fungi, Planta and Animalia at last. *Gene Ther. Mol. Biol.* 4:313–322.

E. Trifonov et al, 2001. Distinct Stages of Protein Evolution as Suggested by Protein Sequence Analysis. *J. Mol. Evol.* 53:394–401.

Enfoques más teóricos

V. Knasnicka, J. Pospichal, T. Kaláb. A Study of Replicators and Hypercycles by Typogenetics. *Lect. Notes Artif. Int.* 2159:37–54.

P. Cisek, 1995. Theoretical Approaches to the Origin of Life. *Internal report at some university...*

http://cns-web.bu.edu/pub/pavel/papers/origin.ps.gz

G. Jenkins, 2001. Evidence for the Non-quasispecies Evolution of RNA Viruses. *Mol. Biol. Evol.* 18:987–994.

E. Szathmáry, 1997. The first two billion years. *Nature* 387:662–663.

E. Szathmáry, J. Maynard Smith, 1997. From Replicators to Reproducers: the First Major Transitions Leading to Life. *J. Theor. Biol.* 187:555–571.

E. Szathmáry, 2000. Reply: certain uncertainties about the origin of the genetic code [Letter]. *Trends Genet.* 16:18–19.

I. Scheuring, E. Szathmáry, 2001. Survival of Replicators with Parabolic Growth Tendency and Exponential Decay. *J. Theor. Biol.* 212:99–105.

S. Lifson, 1997. On the Crucial Stages in the Origin of Animate Matter. *J. Mol. Evol.* 44:1–8.

S. Lifson, H. Lifson, 1999. A Model of Prebiotic Replication: Survival of the Fittest versus Extinction of the Unfittest. *J. Theor. Biol.* 199:425–433.

S. Lifson, H. Lifson, 2001. Coexistence and Darwinian Selection among Replicators: Response to the Preceding Paper by Scheuring and Szathmary. *J. Theor. Biol.* 212:107–109.

V. Kunin, 2000. A system of two polymerases - A model for the origin of life. *Origins Life Evol. B.* 30:459–466.

M. Yčas, 1999. Codons and hypercycles. *Origins Life Evol. B.* 29:95–108.

J. Ricard et al, 1996. The Origin of Kinetic Cooperativity in Prebiotic Catalysts. *J. Mol. Evol.* 43:315–325.

A. Kanavarioti, 1998. Preference for Internucleotide Linkages as a Function of the Number of Constituents in a Mixture. *J. Mol. Evol.* 46:622–632.

L. Monteiro, J. Piqueira, 1998. A Model for the Early Evolution of Self-replicating Polymers. *J. Theor. Biol.* 191:237–248.

L. Monteiro, J. Piqueira, 1998. Numbering Self-replicating Polymers. *J. Theor. Biol.* 193:365–367.

L. Monteiro, J. Piqueira, 1998. Numbering Self-replicating Polymers II. *J. Theor. Biol.* 194:159–161.

L. Monteiro, J. Piqueira, 1999. Modeling Homopolymer Self-replication: Implications for Early Competition. *J. Theor. Biol.* 196:51–60.

F. Mosqueira, S. Ramos-Bernal, A. Negrón-Mendoza, 2002. Biased polymers in the origin of life. *Biosystems* 65:99–103.

R. Happel, P. Stadler, 1998. The Evolution of Diversity in Replicator Networks. *J. Theor. Biol.* 195:329–338.

M. Steel, 2000. The emergence of a self-catalysing structure in abstract origin-of-life models. *Appl. Math. Lett.* 3:91–95.

El último ancestro común y la raíz del Árbol

A. Poole, D. Jeffares, D. Penny, 1998. The path from the RNA world. *J. Mol. Evol.* 46:1–17.

M. Yarus, 1999. Boundaries for an RNA world. *Curr. Opin. Chem. Biol.* 3:260–267.

N. Kyrpides, R. Overbeek, C. Ouzounis, 1999. Universal protein families and the functional content of

the last universal common ancestor. *J. Mol. Evol.* 49:413–423.

D. Penny, A. Poole, 1999. The nature of the last universal common ancestor. *Curr. Opin. Genet. Dev.* 9:672–677.

M. Di Giulio, 2001. The Non-universality of the Genetic Code: the Universal Ancestor was a Progenote. *J. Theor. Biol.* 209:345–349.

M. Di Giulio, 2000. The Universal Ancestor Lived in a Thermophilic or Hyperthermophilic Environment. *J. Theor. Biol.* 203:203–213.

V. Moulton et al, 2000. RNA Folding Argues Against a Hot-Start Origin of Life. *J. Mol. Evol.* 51:416–421

C. Woese, 1998. The universal ancestor. *P. Natl. Acad. Sci. USA* 95:6854–6859.

C. Woese, 2000. Interpreting the universal phylogenetic tree. *P. Natl. Acad. Sci. USA* 97:8392–8396.

J. Castresana, D. Moreira, 1999. Respiratory Chains in the Last Common Ancestor of Living Organisms. *J. Mol. Evol.* 49:453–460.

R. Root-Bernstein, P. Dillon, 1997. Molecular Complementarity I: the Complementarity Theory of the Origin and Evolution of Life. *J. Theor. Biol.* 188:447–479.

P. Dillon, R. Root-Bernstein, 1997. Molecular Complementarity II: Energetic and Vectoral Basis of Biological Homeostasis and its Implications for Death. *J. Theor. Biol.* 188:481–493.

C.P. Ponting et al, 1999. Eukaryotic Signalling Domain Homologues in Archaea and Bacteria. Ancient Ancestry and Horizontal Gene Transfer. *J. Mol. Biol.* 289:729–745.

H. Tiedemann, J. Heitkötter, 1995. Theory of Self-reproducing Automata and Life - Foundation for a Theoretical Biology? *Draft for JTB... never published???.*

S. Kumar, A. Rzhetsky, 1996. Evolutionary Relationships of Eukaryotic Kingdoms. *J. Mol. Evol.* 42:183–193.

P. Forterre, H. Philippe, 1999. Where is the root of the universal tree of life? *BioEssays* 21: 871–879.

Y. Wolf, A. Kondrashov, E. Koonin, 2000. Interkingdom gene fusions. *Genome Biol.* 1:0013.

M. Gouy, W. Li, 1989. Molecular Phylogeny of the Kingdoms Animalia, Plantae, and Fungi. *Mol. Biol. Evol.* 6:109–122.

V. Morell, 1996. Life's Last Domain. *Science* 273:1043–1045.

M. Rivera et al, 1998. Genomic evidence for two functionally distinct gene classes. *P. Natl. Acad. Sci. USA* 95:6239–6244.

M. Belfort, A. Weiner, 1997. Another Bridge between Kingdoms: tRNA Splicing in Archaea and Eukaryotes. *Cell* 89:1003–1006.

K. Rudi, S. Jacobsen, 1997. Cyanobacterial $tRNA^{Leu}$(UAA) group I intron have polyphilogenetic origin. *FEMS Microbiol. Lett.* 156:293–298.

G.B. Golding, R.S. Gupta, 1995. Protein-based Phylogenies Support a Chimeric Origin for the Eukaryotic Genome. *Mol. Biol. Evol.* 12:1–6.

Más sobre tRNA y RNA

P. Veronico, R. Gallerani, L.R. Ceci, 1996. Compilation and classification of higher plant mitochondial tRNA genes. *Nucleic Acid Res.* 24:2199–2203.

F. Damiano et al, 2001. PLMItRNA, a database for mitochondrial tRNA genes and tRNAs in photosynthetic eukaroytes. *Nucleic Acid Res.* 29:167–168.
Database at `http://bio-www.ba.cnr.it:8000/PLMItRNA`.

U. Nagaswamy et al, 2002. NCIR: a database of non-canonical interactions in known RNA structures. *Nucleic Acid Res.* 30:395–397.
Database at `http://prion.bchs.uh.edu/bp_type/`.

B. Knudsen et al, 2001. tmRDB (tmRNA database). *Nucleic Acid Res.* 29:171–172.
Database at `http://psyche.uthct.edu/dbs/tmRDB/tmRDB.html`.

C. Zwieb, I. Wower, J. Wower, 1999. Comparative sequence analysis of tmRNA. *Nucleic Acid Res.* 27:2063–2071.

B. Felden et al, 2001. Phylogenetic analysis of tmRNA genes within a bacterial subgroup reveals a specific structural signal. *Nucleic Acid Res.* 29:1602–1607.

K. Hanawa-Suetsugu et al, 2001. Importance of the conserved nucleotides around the tRNA-like structure of *Eschericia coli* transfer-messenger RNA for protein tagging. *Nucleic Acid Res.* 29:4663–4673.

B. Clouet dÓrval et al, 2001. Box C/D RNA guides for the ribose methylation of archaeal tRNAs. The tRNAtrp intron guides the formation of two ribose-methylated nucleosides in the mature tRNAtrp. *Nucleic Acid Res.* 29:4518–4529.

J. Parsh, H.M. Braverman, W. Stephan, 2000. Comparative Sequence Analysis and Patterns of Covariation in RNA Secondary Structures. *Genetics* 154:909–921.

M. Kinouchi, Y. Kudo, 2001. Detection of tRNA Genes with Introns from DNA Sequences of Archaea. *Genome Informatics* 12:431–432.

M. Helm et al, 2000. Search for characteristic structural features of mammalian mitochondrial tRNAs. *RNA* 6:1356–1379.

R. Giege, M. Sissler, C. Florentz, 1998. Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acid Res.* 26:5017–5035.

L. Marechal-Drouard et at, 1990. Transfer RNAs of potato (Solanum tuberosum) mitochondria have different genetic origins. *Nucleic Acid Res.* 18:3689–3696.

J. Sagara, K. Shimizu, 1999. Analysis of Molecular Evolution in Mitochondrial tRNA Gene Sequences. *Genome Informatics* 10:318–319.

T. Ikemura, 1985. Codon Usage and tRNA Content in Unicellular and Multicellular Organisms. *Mol. Biol. Evol.* 2:13–34.

A. Brennicke et al, 1999. RNA editing. *FEMS Microbiol. Rev.* 23:297–316.

P. Schuster, W. Fontana, 1999. Chance and necessity in evolution: lessons from RNA. *Physica D* 133:427–452.

S. Kopp, C. Reidys, P. Schuster, 1997. Insights into Evolution of RNA-Structures. In *Fourth European Conference on Artificial Life*, MIT Press, 1997.

B. Ma, L. Wang, K. Zhang, 2002. Computing similarity between RNA structures. *Theor. Comput. Sci.* 276:111–132.

F. Xu et al, 2001. Species-specific differences in the operational RNA code for aminoacylation of tRNAtrp. *Nucleic Acid Res.* 29:4125–4133.

A. Nakaya, A. Yonezawa, K. Yamamoto, 1996. Classification of RNA Secondary Structures Using the

Techniques of Cluster Analysis. *J. Theor. Biol.* 183:105–117.

Arquès, y comentarios

D. Arquès, C. Michel, 1987. Periodicities in introns. *Nucleic Acid Res.* 15:7581–7592.

D. Arquès, C. Michel, 1987. A Purine-Pyrimidine Motif Verifying an identical Presence in Almost All Gene Taxonomic Groups. *J. Theor. Biol.* 128:457–461.

D. Arquès, C. Michel, 1987. Study of a Perturbation in the Coding Periodicity. *Math. Biosci.* 86:1–14.

D. Arquès, C. Michel, 1990. Periodicities in Coding and Noncoding Regions of the Genes. *J. Theor. Biol.* 143:307–318.

D. Arquès, C. Michel, 1990. A Model of DNA Sequence Evolution. *B. Math. Biol.* 52:741–772.

D. Arquès, C. Michel, 1992. Analysis of Gene Evolution: the software AGE. *CABIOS* 8:5–14.

D. Arquès, C. Michel, 1995. Identification and Simulation of Shifted Periodicities Common to Protein Coding Genes of Eukaryotes, Prokaryotes and Viruses. *J. Theor. Biol.* 172:279–291.

A.J. Koch, J. Lehmann, 1997. About a Symmetry of the Genetic Code. *J. Theor. Biol.* 189:171–174.

Teorías sin polímeros

D. Segré, D. Lancet, 2000. Composing life. *EMBO Rep.* 11:217–222.

D. Segré et al, 2001. The Molecular Roots of Compositional Inheritance. *J. Theor. Biol.* 213:481–491.

T. Schmitt, N. Lehman, 1999. Non-unity molecular hereditability demonstrated by continous evolution *in vitro*. *Chemistry Biol.* 6:857–869.

S. Sowerby, W. Heckl, G. Petersen, 1996. Chiral Symmetry Breaking During the Self-Assembly of Monolayers from Achiral Purine Molecules. *J. Mol. Evol.* 43:419–424.

S. Sowerby, P. Stockwell, W. Heckl, G. Petersen, 2000. Self-programmable, self-assembling two-dimensional genetic matter. *Origins Life Evol. B.* 30:81–99.

S. Sowerby, C. Cohn, W. Heckl, N. Holm, 2001. Differential adsorption of nucleic acid bases: Relevance to the origin of life. *P. Natl. Acad. Sci. USA* 98:820–822.

W. Heckl, 2002. Molecular Self-Assembly and the Origin of Life. *In Astrobiology. The Quest for the Conditions of Life, G. Horneck, C. Baumstarck-Khan Eds. Springer Verlag.*, 361–372.

H. Morowitz et al, 2000. The origin of intermediary metabolism. *P. Natl. Acad. Sci. USA* 97:7704–7708.

A. Lazcano, S. Miller, 1999. On the Origin of the Metabolic Pathways. *J. Mol. Evol.* 49:424–431.

G. Wächtershäuser, 1990. Evolution of the first metabolic cycles. *P. Natl. Acad. Sci. USA* 87:200–204.

E. Blöchl, M. Keller, G. Wächtershäuser, K.O. Stetter, 1992. Reactions depending on iron sulfide and linking geochemistry with biochemistry. *P. Natl. Acad. Sci. USA* 89:8117–8120.

C. de Duve, S. Miller, 1991. Two-dimensional life? *P. Natl. Acad. Sci. USA* 88:10014–10017.

G. Wächtershäuser, 1997. The Origin of Life and its Methodological Challenge. *J. Theor. Biol.* 187:483–494.

G. Wächtershäuser, 2000. Life as We Don't Know It. *Science* 289:1307–1308.

L. Orgel, 2000. Self-organizing biochemical cycles. *P. Natl. Acad. Sci. USA* 97:12503–12507.

N. Hud, F. Anet, 2000. Intercalation-Mediated Synthesis and Replication: A New Approach to the Origin of Life. *J. Theor. Biol.* 205:543–562.

Teorías "no estándar"

V. scCherbak, 2003. Arithmetic inside the universal genetic code. *Biosystems* 70:187–209.

P. Béland, T.H. Allen, 1994. The Origin and Evolution of the Genetic Code. *J. Theor. Biol.* 170:359–365.

V.A. Karasev, V.E. Stefanov, 2001. Topological Nature of the Genetic Code. *J. Theor. Biol.* 209:303–317.

G. Martínez-Meckler et al, 1999. Primitive molecular machine scenario for the origin of the three base codon composition. *Origins Life Evol. B.* 29:203–214.

J. Bashford, L. Tsohantjis, P.D. Jarvis, 1997. Codon and nucleotide assignments in a supersymmetric model of the genetic code. *Phys. Lett. A* 233:481–488.

J. Bashford, L. Tsohantjis, P.D. Jarvis, 1998. A Supersymmetric Model for the Evolution of the Genetic Code. *P. Natl. Acad. Sci. USA* 95:987–992.

P.D. Jarvis and J. Bashford, 1998. Systematics of the genetic code and anticode: history, supersymmetry, degeneracy and periodicity. *In Proceedings of the XII International Colloquium on Group Theoretical Methods in Physics, S.P. Corney, R. Delbourgo and P.D. Jarvis (Eds.) (International Press, Boston) 147-151.*

J. Bashford and P.D. Jarvis, 2000. The Genetic Code as a Periodic Table: Algebraic Aspects. *To appear in Biosystems*; `arxiv.org/abs/physics/0001066`.

K. Ohnishi, 1998. Poly-tRNA-mediated Emergence of mRNAs and Genetic Codes: The Evolutionary Significance of the Poly-tRNA Structure *Third International Conference on Emergence, Helsinki, Finland, 1998.*

J.E. Hornos, Y.M.M. Hornos, 1993. Algebraic model for the evolution of the genetic code. *Phys. Rev. Lett.* 71:4401–4404.

J.E. Hornos, Y.M.M. Hornos, 1994. A search for symmetries in the genetic code. *J. Biol. Phys.* 20:289–294.

J.E. Hornos, Y.M.M. Hornos, M. Forger, 1999. Symmetry and symmetry breaking: an algebraic approach to the genetic code. *Int. J. Mod. Phys. B* 13:2795–2785.

R.D. Kent, M. Schlesinger, B.G. Wybourne, 1998. On algebraic approaches to the genetic code. *Can. J. Phys.* 76:445–452.

L. Frappat, A. Sciarrino, P. Sorba, 2000. Prediction of Physical-chemical properties of amino acids from the genetic code. *To appear in J. Biol. Phys.*; available at `http://xxx.lanl.gov/abs/physics/0007034`.

L. Frappat, A. Sciarrino, P.Sorba, 2001. Crystalizing the genetic code. *J. Biol. Phys.* 17:1.

A. Sciarrino, 2001. A Mathematical Model Accounting for the Organization in Multiplets of the Genetic Code. arxiv:math-ph/0102022.

M. Forger, S. Sachse, 2000. Lie Superalgebras and the Multiplet Structure of the Genetic Code. I. Codon Representations. *J. Math. Phys.* 41:5407–5422.

M. Forger, S. Sachse, 2000. Lie Superalgebras and the Multiplet Structure of the Genetic Code. II. Branching Schemes. *J. Math. Phys.* 41:5423–5444.

M. Jimenez-Montano, C. de la Mora-Basanez, T. Poeschel, 1994. On the Hypercube Structure of the Genetic Code. *In Proc. 3. Int. Conf. on Bioinformatics and Genome Research H. Lim and C. Cantor eds., World Scientific; Page 445.*. Also arxiv:cond-mat/0204044.

M. Jimenez-Montano, C. de la Mora-Basanez, T. Poeschel, 1996. The Hypercube Structure of the Genetic Code Explains Conservative and Non-Conservative Aminoacid Substitutions in vivo and in vitro. *Biosystems* 39:117–125.

M. Altaisky, F. Filatov, 2001. Genetic Information and Quantum Gas. arxiv:quant-ph/0106123.

J.-L. Tyran, 1998. A Two Dimensional Virtual World to Explain the Genetic Code Structure? *Lect. Notes Artif. Int.* 1434:186–192.

Textos generales

M. Nirenberg, 1968. The genetic code. *Nobel Lecture.*

B. Hayes, 1998. The Invention of the Genetic Code. *Am. Sci.* 86:8–14.

L. Orgel, 1998. The origin of life - How long did it take? *Origins Life Evol. B.* 28:91–96.

H. Baltscheffsky et al, 1997. On the Origin and Evolution of Life: An Introduction. *J. Theor. Biol.* 187:453–459.

Vida en la Tierra primitiva

J. McClendon, 1999. The origin of life. *Earth-Sci. Rev.* 47:71–93.

M. van Zuilen, A. Lepland, G. Arrhenius, 2002. Reassessing the evidence for the earliest traces of life. *Nature* 418:627–630.

M. Robertson, M. Levy, S. Miller, 1996. Prebiotic Synthesis of Diaminopyrimidine and Thiocytosine. *J. Mol. Evol.* 43:543–550.

M. Levy, S. Miller, 1999. The Prebiotic Synthesis of Modified Purines and Their Potential Role in the RNA World. *J. Mol. Evol.* 48:631–637.