UNIVERSIDAD NACIONAL
AUTONOMA DE MÉXICO
FACULTAD DE CIENCIAS

Análisis no lineal y matrices de correlación aplicados al estudio de las secuencias genéticas del virus del SIDA

TESIS

que para obtener el título de

F Í S I C O

presenta

Maximino Aldana González

Tesis dirigida por:

Dr. Germinal Cocho Gil

México, D.F. Febrero de 1995

UNIVERSIDAD NACIONAL AUTÓNOMA DE MEXICO

M. EN C. VIRGINIA ABRIN BATULE
Jefe de la División de Estudios Profesionales
Facultad de Ciencias
P r e s e n t e

Los abajo firmantes, comunicamos a Usted, que habiendo revisado el trabajo de Tesis que realiz(ó)ron el pasante(s) Maximino Aldana González

con número de cuenta 8503349-6 con el Título:

Análisis no lineal y matrices de correlación aplicados al estudio de las secuencias genéticas del virus del SIDA

Otorgamos nuestro **Voto Aprobatorio** y consideramos que a la brevedad deberá presentar su Examen Profesional para obtener el título de Físico

GRADO	NOMBRE(S)	APELLIDOS COMPLETOS	FIRMA
Dr.	Germinal	Cocho Gil	
Director de Tesis			
Dr.	Antonio	Quiroz-Gutiérrez	
M. en C.	Raúl	Rechtman Schrenzel	
Dr.	François	Leyvraz Waltz	
Suplente			
Dr.	Rubén	Santamaría Ortíz	
Suplente			

Indice

Agradecimientos ..... *ii*

Introducción ..... *iv*

Capítulo 1: Antecedentes Biológicos 1
- Las moléculas biológicas 1
- La molécula de la vida: el ADN 3
- La molécula mensajera: el ARN 6
- Las proteínas y el código genético 10
- Los ribosomas: "máquinas" sintetizadoras de proteínas 16
- El Dogma Central de la Biología Molecular hasta antes de 1970 19

Capítulo 2: El virus del SIDA 21
- Genes 21
- Material genético no codificador 22
- Virus y retrovirus 26
- El virus del SIDA 28
- El Dogma Central de la Biología Molecular en la actualidad 33

Capítulo 3: El ADN y periodo 3 35
- La degeneración del código 35
- Representación binaria del ADN 38
- Mínimos de energía en cadenas binarias 42
- Mínimos de energía en las secuencias genéticas 47
- El periodo 3 en el origen de la vida 51

Capítulo 4: Matrices de correlación y Mapas de correlación genética 54
- Variabilidad del VIH 54
- Matriz de correlación 55
- Mapas de correlación genética 59
- El origen de la parte imaginaria de los eigenvalores 69

Conclusiones 72

Apéndice 74

Referencias 83

A Juanita, Andrés, Claudia
y mis padres. Porque ellos
han hecho que mi vida sea mucho
más que una mera existencia.

Agradecimientos

En estas líneas quiero expresar mi agradecimiento a todas aquellas personas que me han ayudado en mi formación académica y en particular, en la realización de este trabajo. Pido disulpas de antemano si es que hago alguna omisión, pero han sido tantos los amigos con los que he contado, que necesitaría una gran cantidad de espacio tan sólo para mencionarlos.

En primer lugar, quiero agradecer al Dr. Germinal Cocho Gil quien, como director de este trabajo, ha mostrado una paciencia infinita para conmigo. No solamente ha estado siempre dispuesto a contestar todas mis preguntas, también me ha dado *todo* el apoyo que he necesitado y ha sido, más que un asesor, un gran amigo. Gracias "Germi" por ser el gran hombre que eres y por dejarme aprender de tí.

Agradezco también a mis sinodales Francois Leyvraz, Antonio Quiroz, Rubén Santamaría y Raúl Rechtman por la dedicación y tiempo invertidos en la lectura, sugerencias y correcciones hechas a este trabajo, que muy lejos de ser fastidiosas, me ayudaron a comprenderlo más y a mejorarlo. Gracias sinceraente por su ayuda.

Muy especialmente quiero agradecer al Dr. Antonio Quiroz y al Dr. Gustavo Martínez porque han sido mis maestros y mis amigos, y porque me hacen sentir que mi trabajo es algo importante. Gracias a ambos por ser personas que siempre están dispuestas a ayudar y a compartir todo lo que saben.

Realmente no tengo palabras para expresar mi agradecimiento a mis padres, a mis hermanos y a Juanita. Simplemente sin ellos no hubiera sido posible que yo cursara una carrera profesional. Gracias Mamá y Papá por todo lo que me han dado. Gracias Claudia por tus consejos, que tanto me han ayudado. Gracias Juanita por tu cariño y comprensión y por todas las cosas que hemos compartido. Y muy especialmente, gracias a tí Andrés por ser el niño tan lindo que eres.

Gracias también a todos los amigos del IFUNAM: a Abraham y Fermín por las desveladas de café y Newton, de Einstein y Feynman, "por Boltzmann" que no conoció a Flor, por el Caos, la Cuántica y la Clásica, por la lección que me dieron y por los "tres mosqueteros". A Sergio Mateos (el "galán simpático") por su sincera y valiosa amistad, por las canciones, por Tulum y Cancún, por la comida china, por los cigarros, por la confianza y por el exámen de Cuántica. En verdad, gracias Sergio. Gracias también al "Flaco" que a pesar de estar tan lejos sigue manteniendo la amistad. Gracias a Eduardo por la guitarra y

por las tareas, y a don Humberto por la música y el cumpleaños. Gracias a Darío Moreno por haberme enseñado a Feynman y por las regañadas que tanto me sirvieron para mi formación posterior. Gracias en fin, a todos los "cuates" que de una u otra manera me han ayudado.

Finalmente, agradezco infinitamente y de manera muy especial al Dr. Miguel de Icaza Herrera, quien me abrió las puertas del IFUNAM. Gracias Miguel por todo, por el cubículo, por los seminarios, por el ultrasonido, por mi primer congreso en Acapulco, por el Potzolcalli, por el TeX, por el C, por Ovidio y los Griegos, por Newton, por Mozart y Bethoven, por la Estadística, la Relatividad y F.E.T.I., por las discusiones de Física y de no Física, por el buen vino, por la confianza y por creer en mí, y sobre todo, por la amistad. No está por demás decirlo otra vez: gracias Miguel porque la huella que has dejado en mí no se borrará jamás.

Introducción

¡Ojalá pudiera conocer lo que contiene el mundo en sus entrañas, revelar el desarrollo de las fuerzas activas y la fuente de todas las cosas y abandonar para siempre el juego de las palabras vacías!
Goethe.

En este trabajo aplicamos las técnicas de la Física de Sistemas Complejos y de la Física Estadística al estudio de secuencias genéticas. Nuestro objetivo es, por una parte, dar un criterio de plausibilidad que conlleve a una explicación de la naturaleza del código genético, es decir, de por qué la información contenida en la molécula de ADN se lee de tres bases en tres bases. Por otra parte, queremos desarrollar métodos de análisis que nos revelen cuáles son las estructuras conservadas en las secuencias genéticas, a pesar de las mutaciones que pudieran existir. Estas técnicas de análisis se pueden aplicar no sólo al estudio de secuencias genéticas, sino a cualquier tipo de señal discreta que presente fluctuaciones aleatorias, pero que sin embargo, tenga una estructura general conservada.

Los dos primeros capítulos de este trabajo están dedicados a dar la motivación biológica de nuestros desarrollos ulteriores. En el capítulo 1 hablamos acerca de la estructura molecular de los ácidos nucleicos, ADN y ARN, y de cómo se guarda la información genética en estas moléculas. En el capítulo 2 describimos con algún detalle al virus que causa el SIDA y su mecanismo de infección en las células linfáticas. Con estos capítulos nos proponemos no solamente exponer la motivación biológica de nuestro trabajo, como ya se ha dicho, sino también el tener una referencia "rápida" en donde un estudiante que quiera ingresar a nuestro grupo de investigación, pueda encontrar los conceptos biológicos básicos necesarios para atacar problemas de tipo biológico utilizando las técnicas de la Física de Sistemas Complejos, sin que tenga que consultar, desde el mero principio, los libros de Biología Molecular existentes en la literatura, que aunque algunos son excelentes, se caracterizan también por ser muy extensos.

El capítulo 3 lo dedicamos a contestar una de las preguntas más importantes de la Biología Molecular, hasta ahora no resuelta, a saber, ¿por qué la información genética contenida en el ARN se lee de tres bases en tres bases? Partimos haciendo una representación binaria de una secuencia genética bajo el criterio de que las bases A y T se ligan complementariamente utilizando dos puentes de hidrógeno, mientras que las bases C y G lo

hacen a través de tres puentes de hidrógeno. Por lo tanto, a las bases A y T las llamamos débiles y las representamos por un 0, mientras que a las bases C y G las llamamos fuertes y las representamos por un 1. Tomando la transformada de Fourier discreta de la representación binaria de secuencias genéticas reales tomadas de diferentes VIH, mostramos explícitamente la presencia de un pico muy agudo en el espectro de Fourier precisamente en el periodo 3. Esto no ocurre, por ejemplo, con secuencias binarias generadas al azar, en donde el espectro de Fourier de tales secuencias es completamente irregular.

Para explicar este periodo 3 en las secuencias genéticas, nos fijamos en las energías de interacción, o energías de amarre entre bases consecutivas en la secuencia. Estas energías de diadas (parejas de bases) ya están reportadas en la literatura desde 1986 por los trabajos de Breslauer y Freir *et al*. Sin embargo, estas energías de amarre son un *promedio* de las interacciones de todos los átomos que conforman a las bases. Consecuentemente, lo que hacemos en nuestro análisis es estudiar el comportamiento del sistema fijándonos sólo en sus propiedades globales promedio, y no en los detalles puntuales átomo por átomo. Este tipo de análisis es muy semejante al que hacen las teorías de campo promedio, en donde las fluctuaciones puntuales se desprecian y sólo se consideran las propiedades globales, aquellas para las cuales la longitud de correlación es comparable con las dimensiones del sistema.

Utilizando la tabla de Breslauer–Freir de energías de amarre entre parejas de bases, calculamos entonces la probabilidad de que en una secuencia *aleatoria* aparezca un mínimo de energía. Esta probabilidad la calculamos tanto numéricamente como teóricamente, obteniendo en ambos casos que la distancia promedio entre mínimos consecutivos de energía a lo largo de la secuencia es precisamente 3.

Tomamos ahora en consideración los trabajos de Magnasco *et al.* sobre máquinas moleculares, los cuales concluyen que si una molécula está bajo la acción de un potencial unidimensional cuasiperiódico y asimétrico respecto a uno de sus periodos, y si el sistema está sometido a fluctuaciones aleatorias que no sean del tipo de ruido blanco gaussiano, entonces la molécula comenzará a moverse a lo largo del potencial en una dirección preferencial, siendo las configuraciones más estables aquellas en las que la molécula "descansa" en los mínimos de dicho potencial.

Nuestro trabajo, incorporando los resultados de Magnasco, nos permite concluir que la manera más estable (energéticamente hablando) en la que se puede mover el ARN de transferencia a lo largo del ARN mensajero, es precisamente de tres bases en tres bases, lo cual resulta ser la primera explicación fundamentada (hasta donde sabemos) de la naturaleza del código genético.

Una secuencia genética es un sistema con muchos grados de libertad, ya que es una cadena de miles de bases, y cada base puede tomar cuatro valores. Por tanto, buscar estructuras conservadas en un sistema tal no es una empresa fácil. En el capítulo 4 tomamos las técnicas de la Física Estadística haciendo uso de la matriz de correlación. Esta matriz fue introducida por primera vez en la literatura dentro del marco de los sistemas dinámicos

caóticos, pero se le considera ahora como una medida "natural" de la correlación para sistemas discretos que posean muchos grados de libertad, tales como los que nosotros trabajamos. La parte novedosa de nuestro trabajo es que hemos desarrollado un método gráfico, basándonos en la matriz de correlación, que nos revela muy claramente cuales son las estructuras conservadas que no cambian en las secuencias genéticas de los organismos, pese a la gran cantidad de mutaciones que éstos puedan tener.

El método consiste en diagonalizar a la matriz de correlación $\hat{\mathbf{C}}(d)$, cuyos elementos se definen como la probabilidad de encontrar a la pareja de bases $\alpha\beta$ en la cadena $(\alpha, \beta = A, T, C, G)$ estando la base β separada una distancia d de la base α a lo largo de la cadena (d es la distancia de correlación).

Buscábamos los eigenvalores de la matriz de correlación porque hay un teorema en Mecánica Estadística (teorema de Perron) que asegura que es el eigenvalor más grande de la matriz de correlación del sistema el que más contribuye a la función de partición. Sin embargo, nos topamos con la dificultad de que para ciertas distancias de correlación los eigenvalores eran complejos, y para otras distancias eran reales. Lo que hicimos entonces fue representar a los eigenvalores de la matriz de correlación como "vectores" en el plano complejo. Sumando estos vectores para distancias de correlación que vayan desde $d = 1$ hasta una cierta distancia máxima, obtenemos una representación gráfica de los eigenvalores de la matriz como función de la distancia de correlación. A este tipo de representaciones las hemos llamado *"Mapas de Correlación Genética"*, por nombrarlos de alguna forma.

En estos mapas podemos ver muy claramente las estructuras conservadas en las secuencias genéticas del virus del SIDA, es decir, las estructuras dentro del genoma viral que, pese a las mutaciones, hacen que el virus siga siendo *el* virus del SIDA. Además, estas técnicas de análisis nos sirven para encontrar "parentesco" entre secuencias genéticas pertenecientes a diferentes organismos, tal como lo mostramos con el HIV-2 y el SIV (el SIV es el virus que causa el SIDA en el mono verde africano), así como entre el cúmulo beta de la hemoglobina en el conejo y en el ser humano. Todas las secuencias genéticas fueron tomadas del GENEBANK 94, publicado por Los Alamos National Laboratory.

Cabe mencionar aquí que tanto el trabajo numérico de lo del periodo 3 en las secuencias genéticas, como los cálculos de los eigenvalores de la matriz de correlación y la construcción de los mapas de correlación genética, se llevaron a cabo utilizando programas de computadora codificados en lenguaje **C** desarrollados por el autor.

Para finalizar, terminamos este trabajo con un apéndice en el cual mostramos diferentes Mapas de Correlación Genética pertenecientes a diferentes genes tanto del HIV-1, del HIV-2 y del SIV, como los pertenecientes a la β-globina del ser humano y del conejo. En el apéndice también mostramos los histogramas de la distribución de la distancia entre mínimos de energía consecutivos en las secuencias intergénicas del cúmulo β de la hemoglobina, tanto para el ser humano como para el conejo.

Es muy importante mencionar aquí que pese a que presentamos este trabajo como algo aislado, el el producto de todo un grupo de investigación en el cual, a lo largo del tiempo, han ido formándose y madurando las ideas que ahora presentamos en forma depurada. Consecuentemente, este trabajo no está aislado, sino que está estrechamente relacionado con otros temas de investigación del grupo de Sistemas Complejos del IFUNAM.

Capítulo 1

Antecedentes Biológicos

Demos gracias a Mamá Naturaleza (o a Dios) de que lo que hizo como materia, como materia se comporta.
Antonio Quiroz G.

Las moléculas biológicas.

Hace 150 años, el joven químico alemán Wöhler escribió: "...debo decirles que he podido preparar urea sin la necesidad de tener el riñón de algún animal, ya sea un humano o un perro". Esta frase marcó el final de la concepción acerca de las "fuerzas vitales" propias de los organismos vivos. Ahora sabemos que los organismos vivos no necesitan de ninguna fuerza vital o vitalismo para poder existir. La idea reveladora en los tiempos de Wöhler fue que las criaturas vivas están hechas de los mismos compuestos químicos que la materia inanimada. En nuestro conocimiento actual no hay cabida ya para vitalismos o fuerzas misteriosas, más allá de las leyes de la química y de la física. Esto no quiere decir que no haya misterios en la biología. Por el contrario. Sin embargo, mucho trabajo ha sido hecho y en una cosa estamos de acuerdo: los organismos vivos no solamente siguen las leyes de la química y de la física, sino que además explotan al máximo estas leyes para dar estructuras altamente complejas y organizadas. Comenzaremos este trabajo describiendo las moléculas fundamentales que dan origen a la gran maquinaria celular.

El agua es la sustancia más abundante en las células vivas. Constituye aproximadamente el 70% del peso total de la célula y por ello todas las reacciones químicas intracelulares se llevan a cabo en un medio acuoso. Pero si no tomamos en cuenta al agua, casi todas las moléculas de la célula son compuestos del carbono, elemento cuya valencia es 4. Desde el punto de vista evolutivo, tal vez el carbono fue escogido para ser el componente principal de las moléculas biológicas por su gran versatilidad para formar cuatro enlaces covalentes

o enlaces dobles con otros elementos o con otros átomos de carbono †, dando lugar así a la formación de moléculas muy grandes cuya diversidad en estructura y tamaño es casi ilimitada. Otros elementos químicos de fundamental importancia en las biomoléculas son el oxígeno, el nitrógeno, el hidrógeno, el fósforo y el azufre; sin embargo, el "esqueleto" de las biomoléculas está compuesto de carbono.

En términos generales, las células están compuestas de cuatro tipos diferentes de moléculas orgánicas: los azúcares, los ácidos grasos, los ácidos nucleicos y las proteínas, cada una de las cuales juega un papel fundamental en la estructura y funcionamiento celulares. Los azúcares, por ejemplo, son moléculas cuya función principal es el almacenamiento de energía química que después puede ser utilizada para dirigir alguna reacción metabólica que de otra forma sería irrealizable. Los ácidos grasos también sirven para almacenar energía química, pero su función principal dentro de la célula está en la construcción de membranas (como fosfolípidos).

Sin duda alguna, las proteínas son las moléculas de mayor diversidad y versatilidad dentro de la célula. Las proteínas son las encargadas de regular la actividad, tanto funcional como estructural, de los seres vivos. Dentro de la célula se producen miles de reacciones químicas cada segundo, reacciones que garantizan la vida y la reproducción celular. Estas reaciones tienen que llevarse a cabo de manera organizada, de tal forma que los productos de una reacción sean el substrato de la reacción siguiente. Al conjunto de todas estas reacciones se le conoce con el nombre de *metabolismo celular*. Las proteínas son las encargadas de asegurar que el metabolismo celular se lleve a cabo correctamente, y de hecho, podemos decir sin exagerar que las proteínas *son* el metabolismo celular. En otras palabras, debido a la gran diversidad en tamaños y estructuras de las proteínas, estas moléculas pueden dirigir una reacción química, indicando el comienzo y el final de la reacción. Pueden servir como catalizadores de muy alta eficiencia, o bien como inhibidores de tal manera que la reacción sólo se realice bajo determinadas circunstancias. Nuestro objetivo por el momento no es analizar ni la estructura ni la función de las proteínas, pero es importante señalar que dichas moléculas son las encargadas de asegurar que la "maquinaria celular" funcione bien. Si se compara una célula con una fábrica, podemos hacer la analogía en cuanto a que las proteínas hacen el papel de los trabajadores de la fábrica.

Las células no solamente son "máquinas biológicas" muy complejas y organizadas, sino que además son máquinas que se construyen a sí mismas. Las células se reproducen teniendo descendientes, y la *información del metabolismo celular debe de pasar de la célula madre a las células hijas integramente*. Es decir, una "máquina biológica" tan compleja y ordenada como la célula tiene que tener en algún lado almacenada la *información* de sus características, tanto funcionales como estructurales; y para que la célula pueda cumplir con su principal función, la de reproducirse, esta información debe de poder transmitirse fielmente de la célula madre a las hijas. Los ácidos nucleicos (ADN y ARN) son las moléculas encargadas de almacenar esta información. Como veremos más adelante, en los ácidos nucleicos está almacenada la información, codificada en un alfabeto de cuatro letras,

† Muy raramente en las biomoléculas se tienen enlaces triples del carbono.

de las características funcionales y estructurales de las proteínas, las cuales controlan a su vez el metabolismo celular. Utilizando terminología informática, los ácidos nucleicos representan el "software" –las instrucciones que forman el programa celular del metabolismo, y las proteínas representan el "hardware" –la maquinaria física que ejecuta el programa almacenado en la memoria. †

La molécula de la vida: el ADN.

El "flujo" de la información genética dentro de la célula comienza con el ADN. En esta molécula está almacenada toda la información sobre las caracteísticas funcionales y estructurales de la célula, y yendo más allá aún, en esta molécula se guarda la información de las características de todo el organismo biológico en su conjunto. La estructura molecular del ADN (ácido desoxirribonucleico) fue descubierta en 1953 por J. D. Watson y F. Crick [1,2]. El ADN es una molécula muy larga que en las células eucariotes (células con núcleo) se encuentra dentro del núcleo celular (y en mucho menor cantidad en las mitocondrias). Está compuesta de cuatro monómeros químicos, llamados *nucleótidos* que, al irse enlazando uno seguido de otro, dan lugar a la molécula *lineal* de ADN.

Los nucleótidos a su vez están compuestos de un grupo fosfato, de un azúcar (desoxirribosa) y de una base nitrogenada. Los grupos fosfato y los azúcares son los mismos en todos los nucleótidos, pero hay cuatro tipos diferentes de bases, llamadas adenina (A), guanina (G), timina (T) y citosina (C). En la figura 1.1(a) se muestra la forma general de un nucleótido, y en la figura 1.1b) se muestran las cuatro bases A, T, C y G. Podemos decir que los nucleótidos son los bloques o "ladrillos" que forman a toda la molécula de ADN.

El ADN es una molécula lineal en forma de *doble hélice*, como una escalera de caracol, cuyos "peldaños" están formados por las bases A, T, C o G, y cuyos "barandales" están formados por los azúcares y fosfatos. En una de las cadenas que forman la doble hélice, los nucleótidos se van enlazando utilizando lo que suele llamarse *enlace nucleosídico o enlace fosfodiéster* (mostrado en la figura 1.2(b)): el grupo fosfato de un nucleótido se une covalentemente al átomo de carbono 3'-desoxiribosa del nucleótido siguiente con la liberación de agua. El átomo de carbono 5'-desoxiribosa de este segundo nucleótido tiene ligado covalentemente un grupo trifosfato, el cual a su vez se enlaza con el átomo de carbono 3'-desoxiribosa del siguiente nucleótido con la liberación de agua y de un pirofosfato, que después se disocia en dos fosfatos para completar la reacción, y así sucesivamente.

La otra de las cadenas que forman la doble hélice se construye de manera similar, pero con una peculiaridad. Vemos de la figura 1.1(b) que la adenina (A) solamente se enlaza con la timina (T) por medio de dos puentes de hidrógeno, mientras que la guanina (G)

† Algunas moléculas catalíticas de ARN también pueden clasificarse dentro del "hardware" celular.

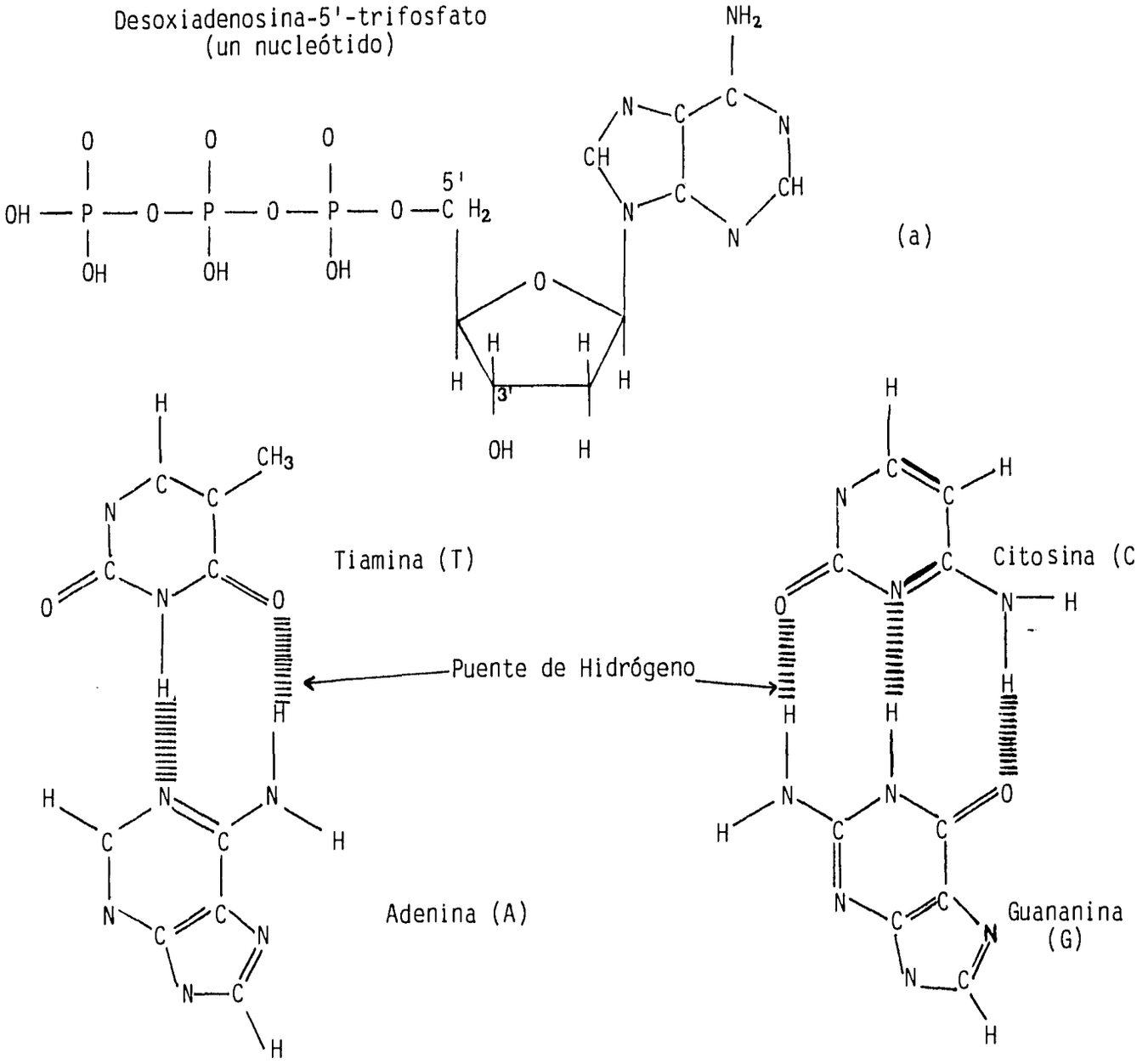


Figura 1.1

(a) Forma general de un nucleótido, el cual consta de una base, un azúcar y un grupo fosfato. (b) Las cuatro bases que pueden estar presentes en los nucleótidos y los enlaces complementarios que se realizan entre ellas. Nótese que la adenina (A) se enlaza con la timina (T) utilizando dos puentes de hidógeno, mientras que la guanina (G) se enlaza con la citosina (C) a través de tres puentes de hidógeno.

sólo se enlaza con la citosina (C) a través de tres puentes de hidrógeno †. Por eso se dice que la adenina es la base *complementaria* de la timina, y que la guanina es la base *complementaria* de la citosina. A los enlaces A–T y C–G se les llama *enlaces complementarios o enlaces de tipo Watson-Crick*. De lo anterior resulta, como muestra la figura 1.2, que las dos cadenas que forman la estructura de doble hélice en el ADN *deben de ser complementarias para que puedan enlazarse*. En otras palabras, si en una de las cadenas se encuentra una A, en la cadena complementaria se debe de encontrar una T en el sitio correspondiente, mientras que si se encuentra una C, en la cadena complementaria se debe de encontrar una G, y viceversa. Este princípio de complementariedad es de fundamental importancia en el almacenamiento y transmisión de la información genética en las células.

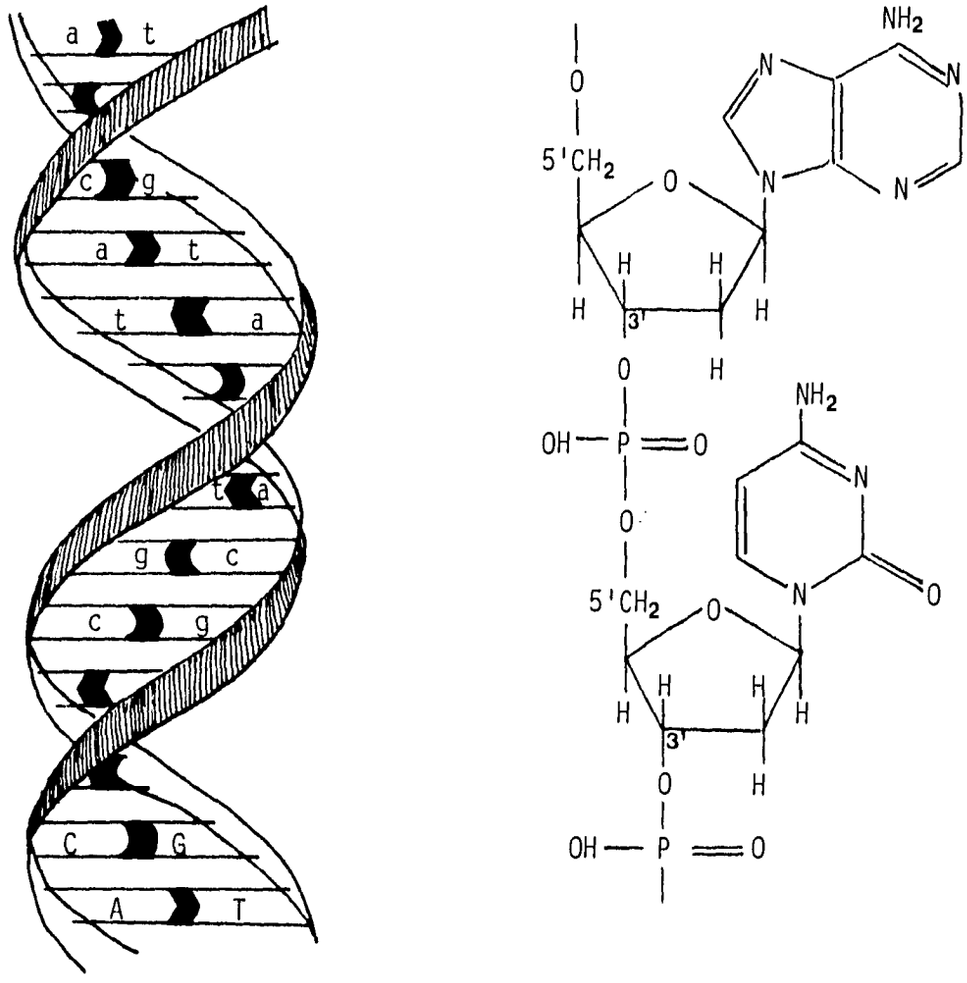


Figura 1.2

(a) Representación esquemática de un pedazo de la molécula de ADN. Nótese la complementariedad de bases que existe entre las dos cadenas que forman a la molécula. (b) Enlace fosfodiéster entre dos nucleótidos.

† Pueden existir también cierto tipo de enlaces poco usuales entre A y C, o entre T y G, pero son mucho más débiles e inestables que los enlaces usuales A–T y C–G.

¿Cómo se guarda la información del metabolismo celular en la cadena de ADN? A través de la secuencia de bases que existe a lo largo de la molécula (de ahora en adelante nos referiremos a las bases químicas adenina, guanina, timina y citosina utilizando las abreviaciones A, G, T y C, respectivamente). Debido a la complementariedad de las dos cadenas, basta con que nos refiramos a una sola de ellas, ya que si en un trozo de una de las cadenas la secuencia de bases es por ejemplo

$$5' \longrightarrow ATTCTGCTA \longrightarrow 3'$$

sabemos que en el trozo correspondiente de la cadena complementaria la secuencia de bases *deberá* de ser

$$3' \longleftarrow TAAGACGAT \longleftarrow 5'$$

Desde el punto de vista de la información contenida en la molécula de ADN, dicha información está codificada en un código de cuatro letras y almacenada en la *secuencia* de bases que existe a lo largo de la molécula de ADN. Es precisamente la secuencia de las bases lo que determina el tipo de información contenida en la molécula, al igual que la secuencia de letras y signos de puntuación en un libro determina el mensaje que porta dicho libro. La diferencia con el caso del libro es que en el ADN, el "alfabeto" sólo consta de cuatro "letras" y no hay signos de puntuación.

La idea de complementariedad de las cadenas y de la secuenciación de las bases ha sido tan importante en el entendimiento de la biología celular, que es difícil imaginar el gran abismo intelectual que vino a llenar.

La molécula mensajera: el ARN

Como ya hemos dicho, la molécula de ADN se encuentra siempre en el núcleo celular y en pequeñas cantidades en las mitocondrias (en las células eucariotes), mientras que los *ribosomas*, que son los organelos celulares encargados de leer y expresar la información genética contenida en el ADN, se encuentran en el citoplasma, fuera del núcleo celular, y particularmente forman el retículo endoplasmático. Para que la información genética contenida en el ADN pueda llegar desde el interior del núcleo hasta los ribosomas, la célula utiliza una molécula intermediaria llamada ARN (ácido ribonucleico) que se forma en el interior del núcleo y que puede viajar atravesando la membrana nuclear hasta llegar a los ribosomas, *y que transporta la misma información que la molécula de ADN*. De hecho, la molécula de ARN es muy similar, tanto química como estructuralmente, al ADN.

El ARN está formado por cuatro nucleótidos, los cuales a su vez están formados por un azúcar, un grupo fosfato y una base nitrogenada, como se muestra en la figura 1.3. Nótese que el azúcar de los nucleótidos del ARN es una *ribosa* y no una *desoxiribosa*

como en el ADN. El hecho de que el azúcar que forma parte de los ARN–nucleótidos tenga un grupo hidroxilo (–OH) extra en la posición 2', hace que la molécula de ARN sea muy susceptible a la hidrólisis (disolverse en agua), lo cual no ocurre con el ADN, que es una molécua muy estable.

Por otro lado, las bases químicas que forman parte de los nucleótidos del ARN, son las mismas que las del ADN, excepto por que la base timina (T) del ADN ha sido sustituida

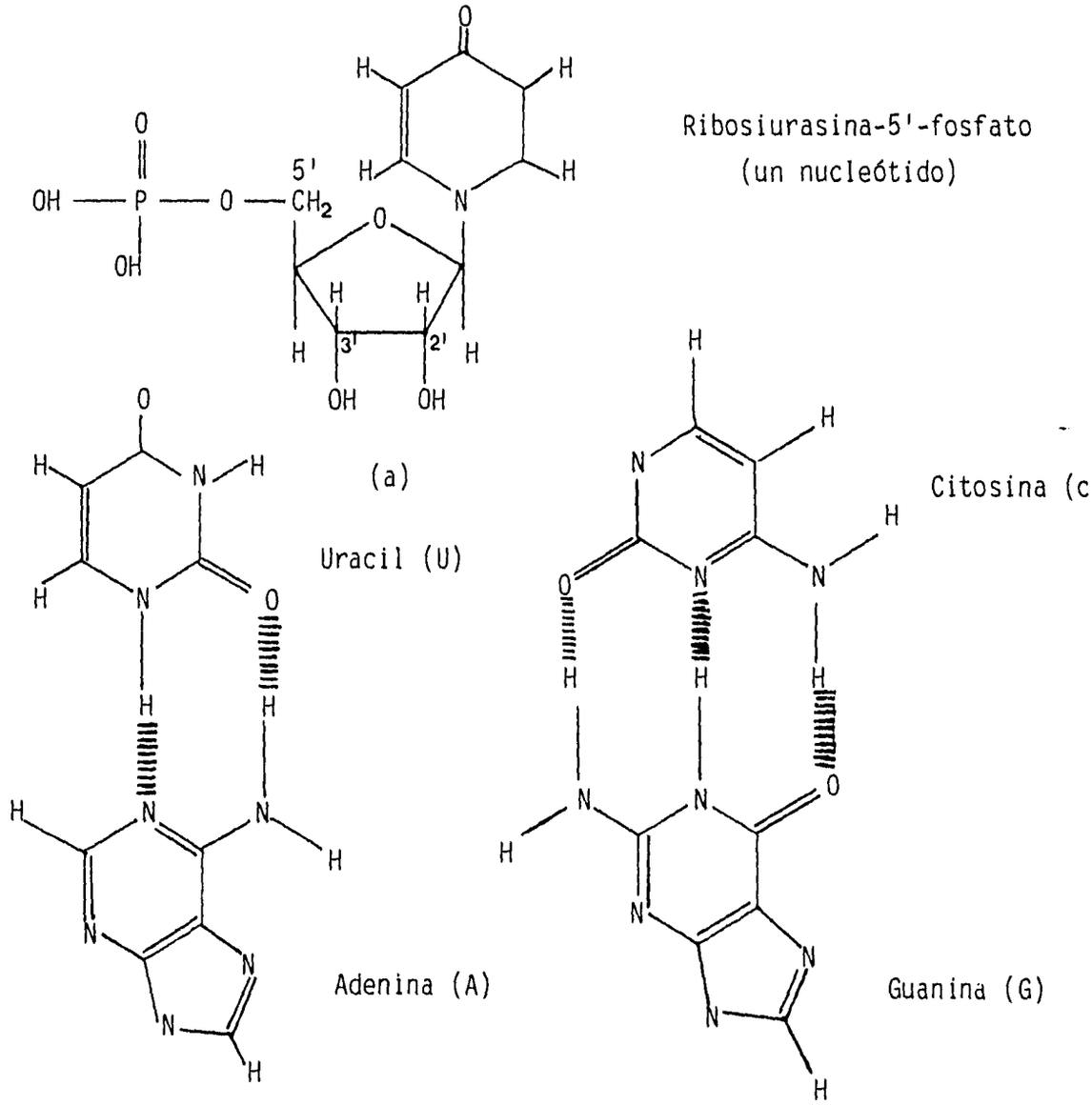


figura 1.3

Los nucleótidos del ARN. (a) Fórmula estructural de un nucleótido típico. (b) Enlaces complementarios que se dan entre las bases del ARN: A–U y C–G.

por otra base muy similar llamada *uracil* (U). Las cuatro bases químicas del ARN son entonces: adenina (A), uracil (U), citosina (C) y guanina (G). Y una característica importante es que también existe un princípio de complementariedad entre estas bases: la A sólo se enlaza con la U a través de dos puentes de hidrógeno, mientras que la C sólo se enlaza con la G utilizando tres puentes de hidrógeno. En todos los sentidos, estas cuatro bases son idénticas desde el punto de vista químico a las cuatro bases del ADN. Podemos decir que el ARN es una molécula que para almacenar la información genética, también utiliza un alfabeto de cuatro letras, la diferencia es que ahora esas letras son A, U, C y G, y no A, T, C y G, como en el ADN. Además, la manera en que se enlazan los ARN-nucleótidos para construir la molécula lineal de ARN es exactamente la misma que la manera en que se enlazan los ADN-nucleótidos del para construir al ADN.

Aunque las moléculas de ADN y ARN son muy similares desde el punto de vista de que los "bloques" fundamentales con los que están construidas son muy similares, existe una diferencia estructural de la mayor importancia entre ambas moléculas: al contrario del ADN, que es una molécula lineal formada por dos cadenas complementarias que dan origen a una estructura de doble hélice, el ARN es una molécula *unicatenaria*, es decir, *está compuesta de una sola cadena*, de modo que los enlaces complementarios entre las bases A–U y C–G se llevan a cabo entre bases que están en diferentes partes de la *misma* cadena, dando como consecuencia que la molécula de ARN se pliegue sobre sí misma, como se muestra en la figura 1.4.

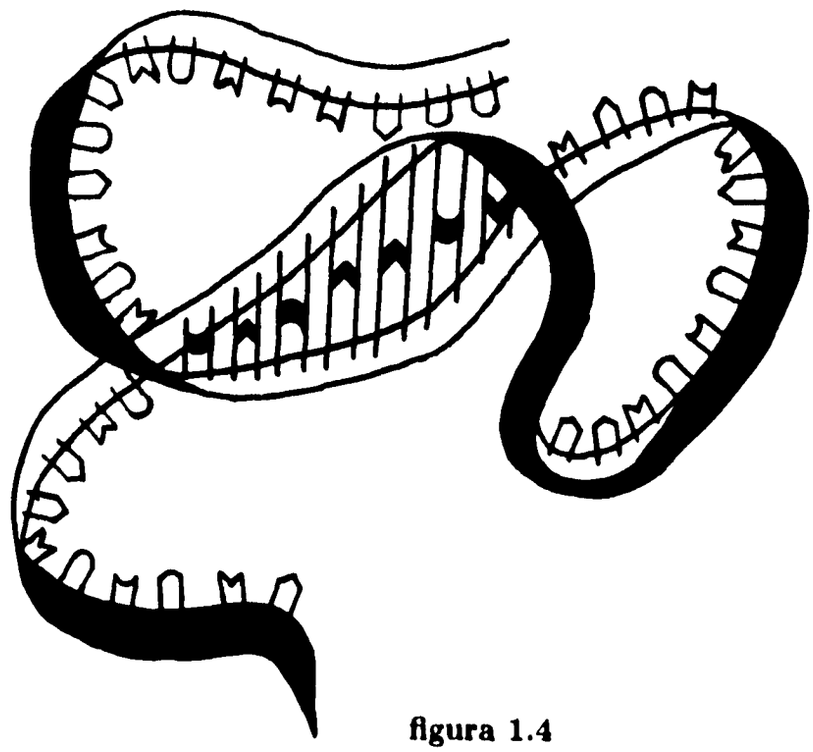


figura 1.4

En la molécula de ARN, los enlaces complementarios A–U y C–G se realizan entre bases que se encuentran en diferentes partes de la misma molécula, dando como consecuencia que ésta se enrolle sobre sí misma

No discutiremos aquí detalladamente cómo se sintetiza la molécula de ARN a partir del ADN dentro del núcleo, pero la figura 1.5 ilustra el proceso, conocido con el nombre de *transcripción*. La doble cadena de ADN se abre, de tal manera que una de las dos cadenas sirve como "molde" para la formación de una cadena *complementaria* de ARN. Una enzima llamada *RNA polimerasa-DNA dirigida*, o a veces simplemente *transcriptasa*, es la encargada de ir añadiendo los nucleótidos correspondientes, uno a la vez, al extremo 3' de la cadena lineal de ARN, basándose en la cadena molde de ADN. Para que la transcriptasa pueda "leer" la secuencia de bases que hay en el ADN mientras está sintetizando la molécula de ARN, las dos cadenas que forman al ADN deben de estar abiertas, ya que en la doble hélice las bases se encuentran "dentro" de los barandales de azúcar y fosfato. Mientras la transcriptasa va recorriendo la molécula de ADN, las dos hebras que forman la doble hélice se van abriendo e inmediatamente después que pasó la transcriptasa, la doble hélice se cierra (como si fuera un cierre de pantalón). Una vez que se ha sintetizado el pedazo de ARN que llevará la información a los ribosomas, la transcriptasa se desliga del ADN y del ARN recién formado, el cual sale del núcleo para ser leído por los ribosomas. El hecho de que la cadena de ARN se forme de manera complementaria a una de las cadenas de ADN, garantiza que el ARN contenga la misma información genética que la molécula precursora de ADN. Como este ARN formado en el núcleo lleva el mensaje genético, se le llama ARN *mensajero*, o simplemente *mARN*.

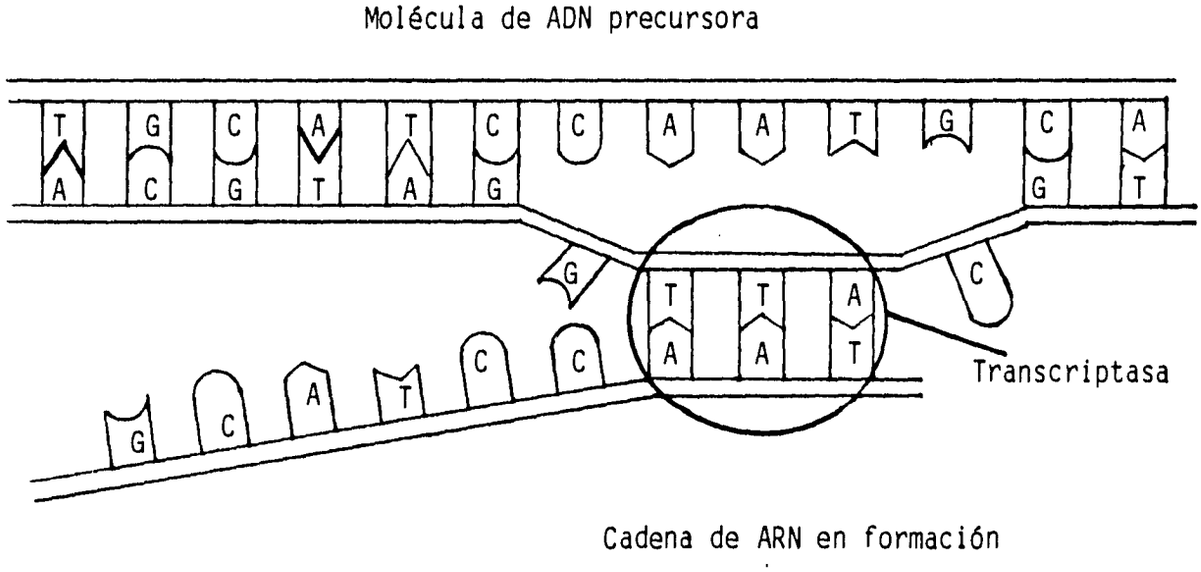


Figura 1.5

Diagrama esquemático del proceso de transcripción que se lleva a cabo dentro del núcleo celular para sintetizar una molécula mensajera de ARN a partir de una de las cadenas de la molécula de ADN. Nótese que la cadena de ARN es complementaria a una de las cadenas de ADN, de modo que contiene la misma información genética que la molécula precursora de ADN.

Las proteínas y el código genético.

Las proteínas son las moléculas encargadas de controlar el metabolismo celular. A pesar de la gran diversidad de proteínas existentes en los organismos vivos, todas las proteínas están formadas por solamente 20 tipos diferentes de moléculas, llamadas *aminoácidos*, que son los mismos en todos los organismos, desde el más "simple" hasta el más "complicado". Al igual que los nucleótidos son los "ladrillos" con los que se construye el ADN, así los aminoácidos son los bloques fundamentales con los que se construyen las proteínas.

Los 20 aminoácidos se muestran en la figura 1.6. Como puede verse de esta figura, todos los aminoácidos tienen un grupo *amino* ($-NH_2$) y un grupo *carboxilo* ($-COOH$), y en lo que se diferencían es en el *radical* (R), que está unido al mismo átomo quiral de carbono al que están unidos los grupos amino y carboxilo. Es precisamente el radical R el que les confiere las propiedades químicas características de cada aminoácido. Por ejemplo, el radical del aminoácido *arginina* es básico, debido a que contiene al grupo ($-NH_2^+$), mientras que el *ácido aspártico* tiene un radical que es ácido debido al grupo carboxilo. Por otro lado, el radical de la *tirosina* es neutro pero polar, porque contiene al grupo hidroxilo ($-OH$), mientras que por la gran estabilidad del anillo de benceno, el radical de la *fenilalanina* es neutro no polar.

Para formar las proteínas, los aminoácidos se van uniendo de forma lineal utilizando el *enlace peptídico*, el cual se forma por eliminación de los elementos de H_2O del grupo carboxilo de uno de los aminoácidos y del grupo α-amino del otro, mediante la acción de agentes de condensación enérgicos. El enlace peptídico se ilustra en la figura 1.7. Una de las principales razones por la cual existe una gran diversidad de proteínas, es porque el enlace peptídico es muy flexible y permite que la molécula lineal que se forma por la unión de varios aminoácidos pueda "enrollarse" sobre si misma, adoptando estructuras que básicamente dependen del tipo de radicales que se tengan a lo largo de la molécula.

Hasta este momento, solamente hemos dicho que la molécula de ADN contiene la información del metabolismo celular. Sin embargo ¿Cuál es exactamente el tipo de información que transporta la molécula de ADN? Además de contener la información de ciertos tipos de ARN's, en el ADN está contenida la información de la *secuencia* de aminoácidos que conforman a las proteínas:

> *la secuencia de nucleótidos en el ADN determina la secuencia de aminoácidos en las proteínas.*

Debido a que el ARN es una molécula complementaria al ADN, y por lo tanto transporta la misma información genética, lo anterior también puede escribirse como "la secuencia de bases en el ARN determina la secuencia de aminoácidos en las proteínas". Sin embargo, hay que tener presente que el mARN es solamente una molécula intermediaria y que la principal molécula almacenadora de la información genética, es el ADN.

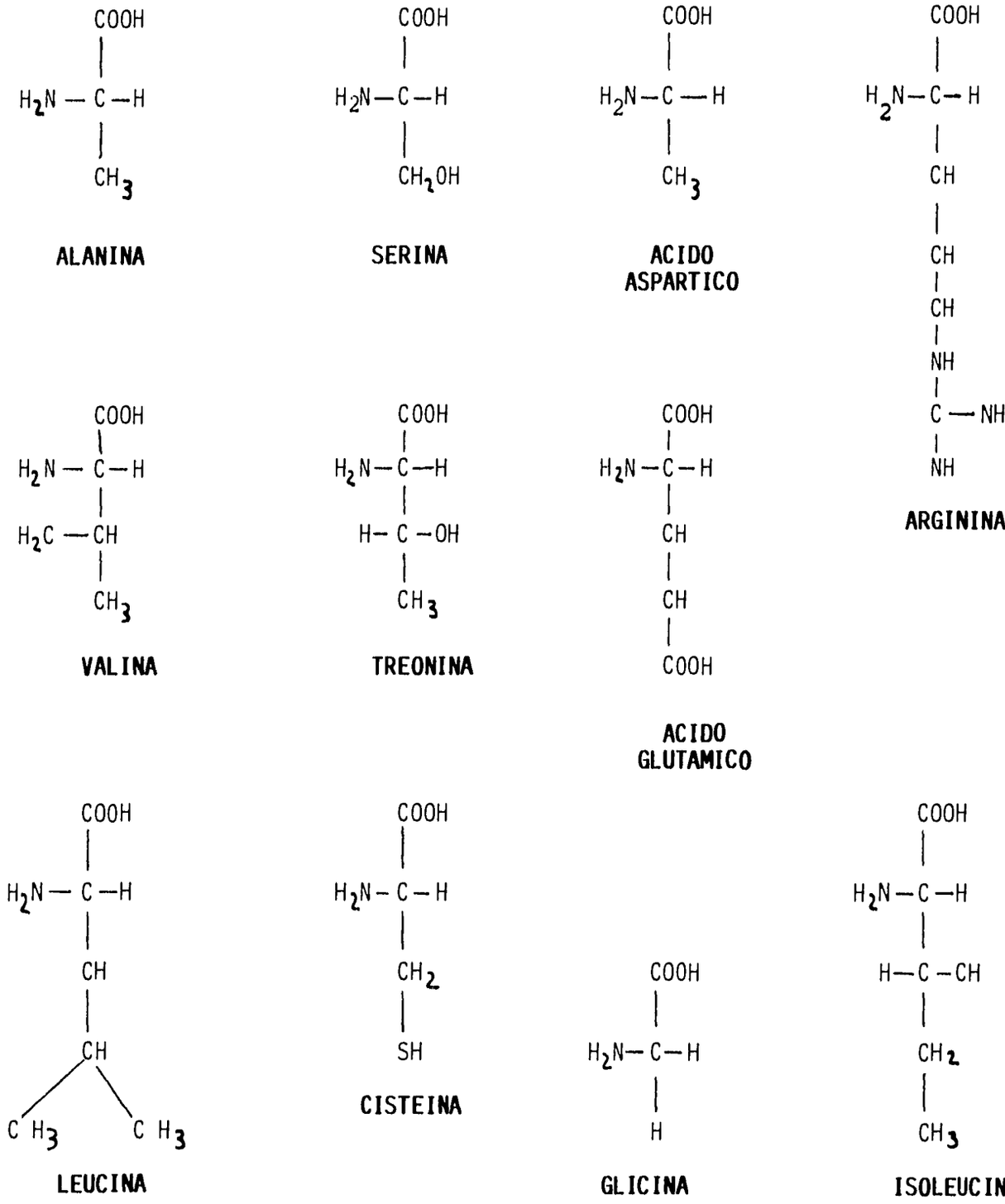


Figura 1.6

Los 20 aminoácidos: las moléculas sillares de las proteínas

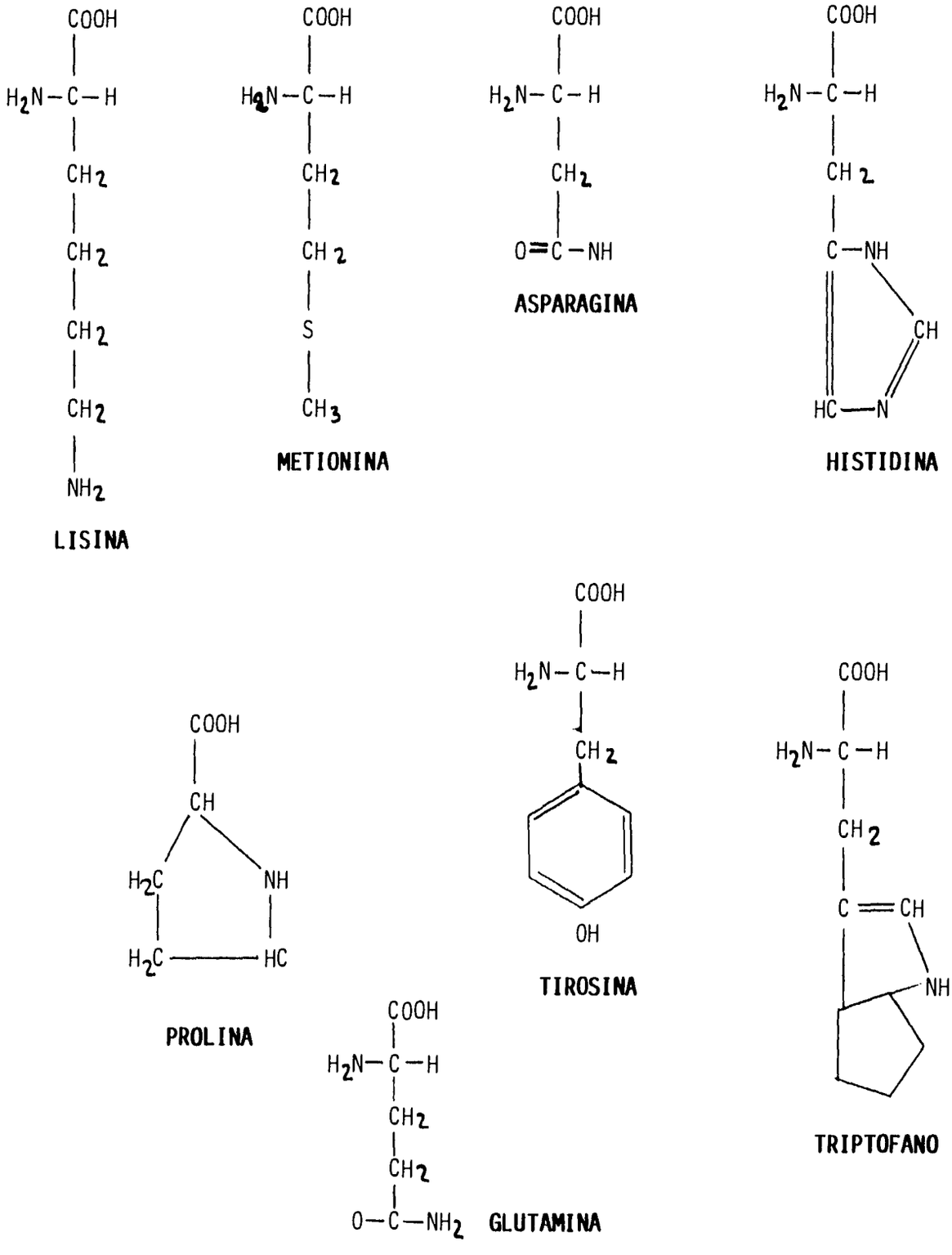
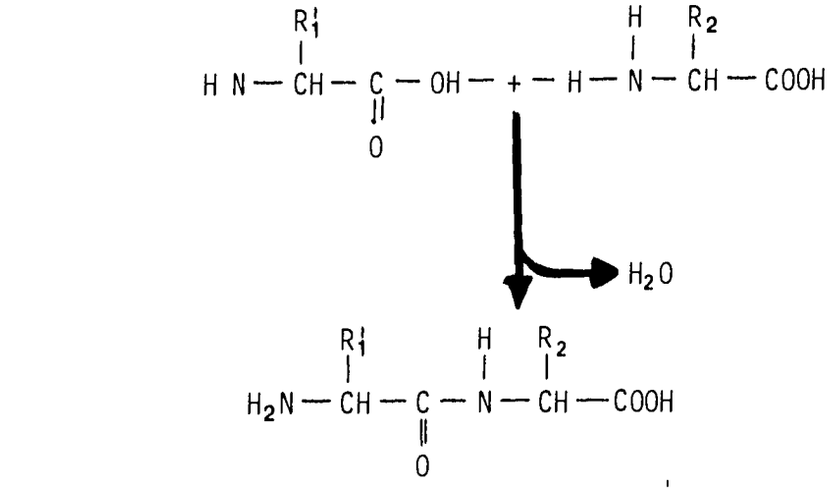
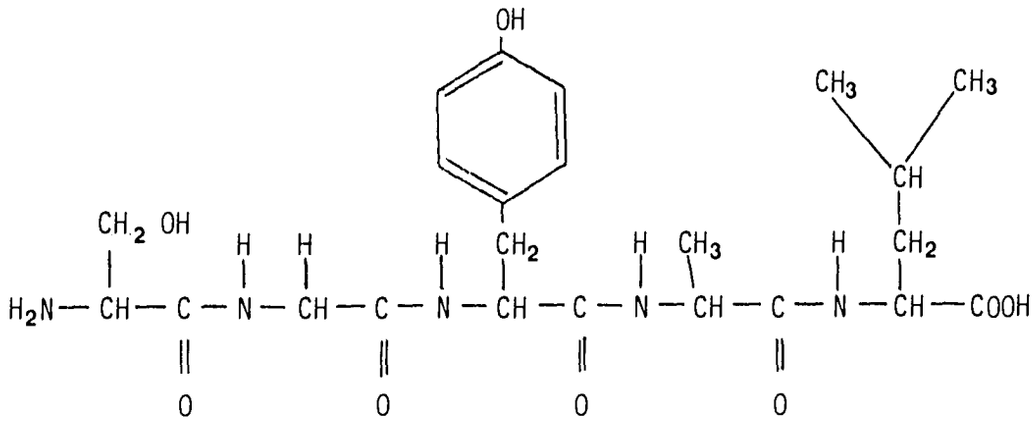


Figura 1.6 (cont.)

Los 20 aminoácidos: las moléculas sillares de las proteínas



(a)



(b)

Figura 1.7

a) El enlace peptídico de los aminoácidos se forma entre el grupo carboxilo de un aminoácido y el grupo amino del siguiente aminoácido, con la liberación de agua. b) En este péptido se muestra la secuencia de aminoácidos Ser–Gly–Tyr–Ala–Leu

Fueron los trabajos pioneros realizados por Marshall Nirenberg *et al.* en la década de los años 60's [3,4] los que dilucidaron lo que después llegó a conocerse como el *código genético* . Si pensamos en las bases químicas A, U, C y G † como un alfabeto de cuatro letras, los "vocablos" que utiliza la célula para codificar la información genética están constituidos por *triadas* de bases. A cada triplete de bases se le llama *codón*, y cada codón "significa", dentro del "lenguaje" genético, un aminoácido. Por ejemplo el codón CCC codifica para el aminoácido prolina, el codón AAA codifica para el aminoácido lisina, mientras que el codón CAU especifica el aminoácido histidina. En la tabla siguiente está el código genético completo.

Código Genético.

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

La idea central en el código genético es que la información se va leyendo, a lo largo de la cadena de ADN, de forma lineal de tres bases en tres bases, o bien, por codones. Por ejemplo, utilizando la tabla del código genético, vemos que la cadena de ADN cuya secuencia de bases es

† Como es el ARN el que está involucrado directamente en la síntesis de proteínas, el código genético se da normalmente en términos de las bases del ARN, y no del ADN.

$5' \rightarrow AUCGAAUUCAUGGUAGCUAGGGCUAGGAAAAAGUCGAUGCCC \rightarrow 3'$

codifica para el péptido cuya secuencia de aminoácidos es

$Ile - Glu - Phe - Met - Val - Ala - Arg - Ala - Arg - Lys - Lys - Ser - Met - Pro$

Notemos que la lectura de la cadena de ADN se lleva a cabo en la dirección $5' \rightarrow 3'$ (véase la figura 1.2), y que *no* hay signos de puntuación en la lectura: los codones se leen uno seguido de otro. Notemos además que el código genético es *degenerado*, es decir, que diferentes codones codifican para el mismo aminoácido. Un ejemplo lo tenemos en el aminoácido alanina (Ala), el cual es codificado por los codones GCU, GCA, GCC y GCG. En el ejemplo anterior, es importante resaltar el hecho de que todos los codones que codifican para alanina, sólo difieren en la última letra. Si observamos detenidamente la tabla del código genético, podremos ver que casi genéricamente los codones que codifican para el mismo aminoácido difieren únicamente en la tercera posición, lo cual nos sugiere que son las dos primeras letras de cada codón las determinantes principales de su especificidad; la tercera posición, es decir, la base situada en el extremo 3' del codón, es menos específica.

Existen tres codones que no especifican a ningún aminoácido: los codones UAA, UAG y UGA tienen la función de indicar el fin de la lectura del mensaje. Por esta razón, a estos codones se les llama *codones de término.* Por otro lado, el codón AUG, además de especificar el aminoácido metionina, indica el *inicio* de la lectura del mensaje. Cualquier hebra de ARN que fuera leída para sintetizar una proteína, tiene que comenzar con el codón de inicio AUG y terminar con alguno de los codones de término. Estos codones son de fundamental importancia, ya que la presencia de los codones de término y de inicio dentro de la cadena del mARN (y por consiguiente, dentro del ADN) garantiza que sólo sea leído el pedazo de material genético necesario para sintetizar una proteína. Llamaremos a este pedazo de material genético que codifica para una proteína completa, un *gen* †. Con esta definición, lo anterior puede expresarse de la siguiente manera: un gen siempre comienza con el codón de inicio AUG y termina con alguno de los codones de término UAA, UAG o UGA.

Otra característica muy importante del código genético es su *universalidad.* Los vocablos del código genético son los mismos en todos los organismos que han sido investigados, que incluyen a los seres humanos, a la bacteria *E. coli*, a la planta del tabaco y a los anfibios, entre otras muchas especies, así como en los genomas virales. De ello parece desprenderse que todas las especies de plantas y animales poseen un precursor evolutivo común, cuyo código genético ha sido completamente protegido a lo largo del curso de la evolución biológica. El desciframiento del código genético ha sido considerado como el descubrimiento científico más importante de la década de los años 60's.

† En el capítulo siguiente daremos una definición más precisa de lo que es un gen.

Los ribosomas: "máquinas" sintetizadoras de proteínas.

Los codones en una molécula de mARN no reconocen directamente el aminoácido que especifican en la forma en que una enzima reconoce a su substrato. El proceso de *traducción*, es decir, la síntesis de una proteína a partir de la molácula de mARN, depende de una molécula "adaptadora" que reconoce a un codón y al aminoácido especificado por ese codón. Esos adaptadores consisten de un conjunto de pequeñas moléculas de ARN conocidas como *ARN de transferencia* o simplemente *tARN*, cada una formada por aproximadamente 80 nucleótidos. A cada aminoácido le corresponde, como mínimo, una clase de tARN; algunos aminoácidos poseen dos o más tARN's específicos. Para reconocer todos los codones de los aminoácidos, son necesarios por lo menos 32 tARN's, y en algunas células, este número es mucho más elevado.

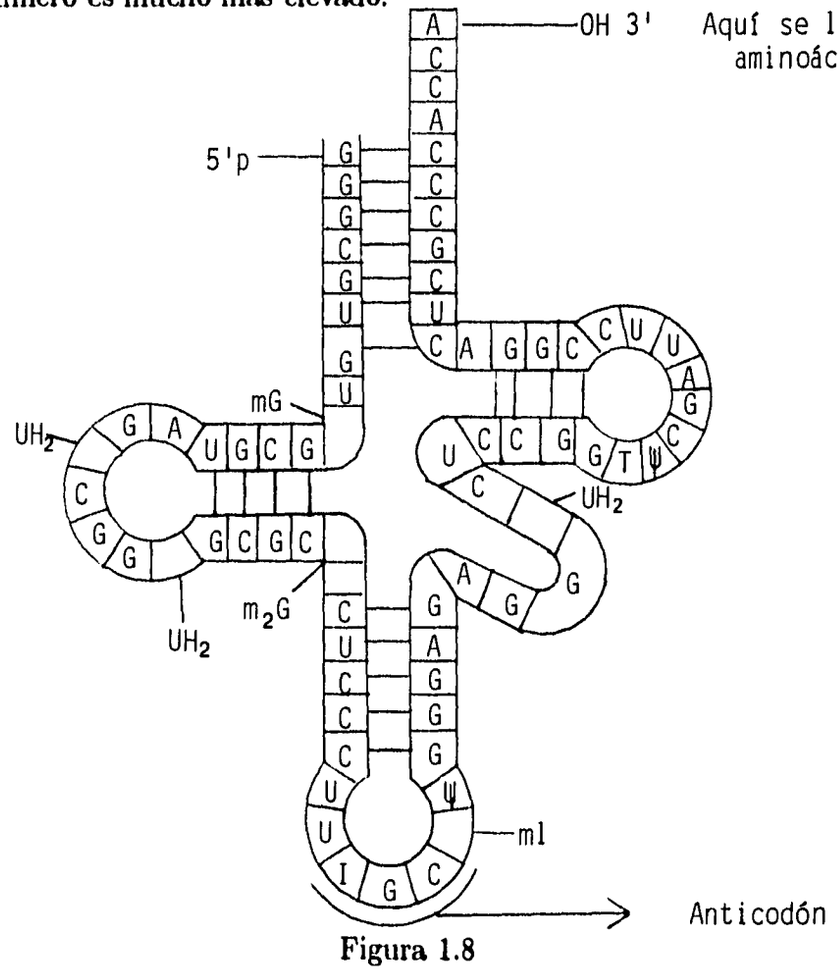


Figura 1.8

Secuencia nucleotídica del tARN de la alanina. En la conformación bidimensional la molécula tiene forma de "trebol". El brazo inferior del trebol tiene el anticodón, un triplete específico de nucleótidos que es complementario y puede aparear sus bases de modo antiparalelo con su triplete codón correspondiente, presente en el mARN. El extremo terminal 3' de la molécula contiene el sitio donde se liga el aminoácido correspondiente al codón reconocido por el brazo anticodón. Nótese que además de las bases usuales A, U, C y G, existen otros residuos nucleotídicos que contienen bases modificadas poco frecuentes, la mayor parte de las cuales son derivados metilados de las bases principales.

La molécula de tARN es unicatenaria, pero enlaces complementarios entre nucleótidos que se encuentran en diferentes partes de la cadena hacen que la molécula se enrolle sobre sí misma, adoptando una estructura tridimensional en forma de "L". En 1965 Robert W. Holley y sus colaboradores de la Universidad de Cornell, obtuvieron la secuencia nucleotídica completa del tARN de la alanina. Este ácido nucléico, que fue el primero en ser secuenciado totalmente, contiene 76 residuos nucleotídicos. En la figura 1.8 se muestra un esquema bidimensional de esta molécula. Todos los tARN's tienen la misma estructura que el mostrado en la figura 1.8. Notemos de la figura que hay dos partes de interés en el tARN: la primera es la parte del *anticodón*, la cual "reconoce" y luego se une al codón complementario en la cadena del mARN. La segunda parte de interés es el extremo 3' terminal, que es el lugar de la molécula donde se une el aminoácido especificado por el codón complementario al anticodón. La figura 1.9 muestra esquemáticamente cómo los tARN's se van uniendo a la molécula de ARN mensajero para formar la proteína.

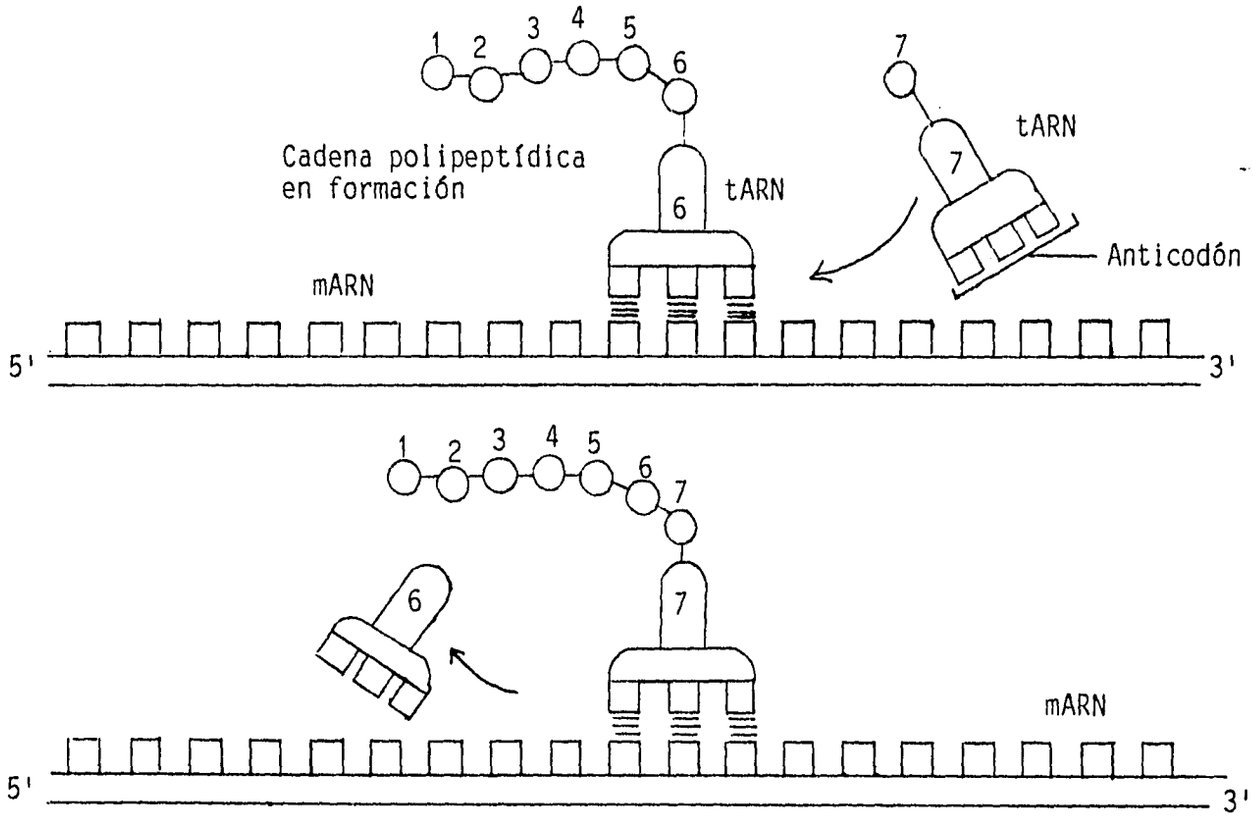


Figura 1.9

El anticodón del tARN se liga de forma complementaria al codón correspondiente en el mARN, mientras que el el otro extremo de la molécula, un aminoácido específico se mantiene unido por un enlace covalente al extremo terminal 3' del tARN. Cuando ocurre el apareamiento entre el codón y el anticodón, este aminoácido se une al extremo α-amino terminal de la cadena polipeptídica en formación. Así, la traducción de la secuencia de nucleótidos en el mARN a la secuencia de aminoácidos en la proteína depende del apareamiento complementario entre el anticodón en el tARN y el correspondiente codón en el mARN

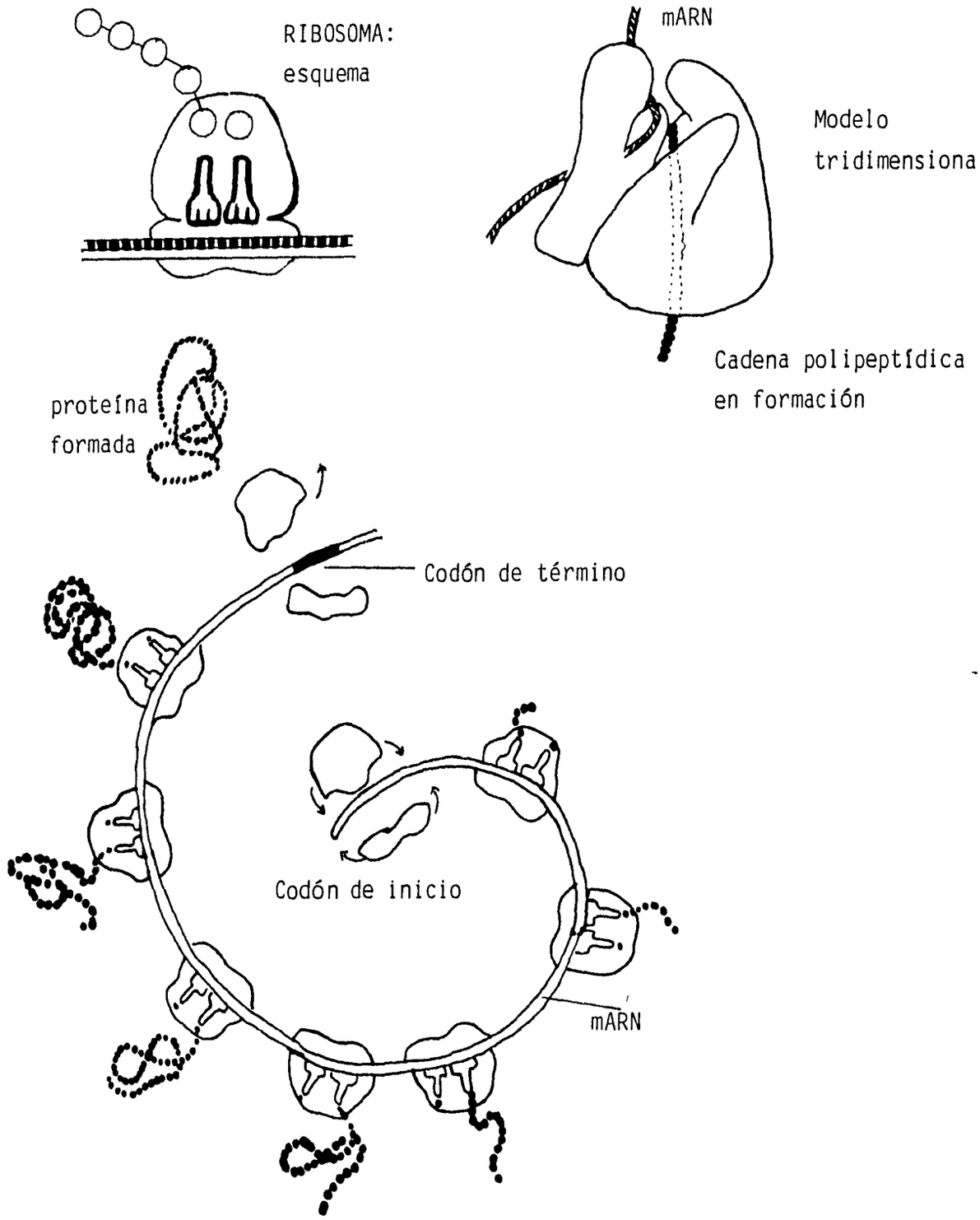


Figura 1.10

Síntesis de proteínas por ribosomas ligados a una molécula de mARN. Los ribosomas se ligan cuando encuentran el codón de inicio cerca del extremo 5' del mARN y luego se mueven hacia el extremo 3' de esta molécula, sintetizando la proteína mientras se van moviendo. El proceso se termina cuando el ribosoma encuentra alguno de los codones de término. Entonces tanto el ribosoma como la proteína se liberan del mARN. Normalmente, la molécula de mARN tiene ligados varios ribosomas al mismo tiempo, cada uno de los cuales está fabricando la misma proteína; la estructura completa se conoce como "polirribosoma".

El proceso de "reconocimiento de codones" a través del cual la información genética se transfiere del mARN a las proteínas, via el tARN, está basado en la formación de enlaces complementarios entre bases. Sin embargo, la manera en cómo se van ordenando los tARN's a lo largo de la cadena de mARN está dirigida por los *ribosomas*, que son moléculas muy grandes compuestas de más de 50 proteínas diferentes asociadas a varias moléculas estructurales de ARN, llamadas *ARN ribosomal o rARN*. Cada ribosoma es una máquina muy grande sintetizadora de proteínas dentro de la cual las moléculas de tARN se acomodan de tal manera que puedan leer el mensaje genético codificado en la molécula de mARN. El ribosoma primero encuentra el codón de inicio en la cadena de mARN, a partir del cual comenzará la lectura. Luego, conforme el ribosoma se mueve a lo largo de la molécula de mARN, traduce la secuencia de nucleótidos en una secuencia de aminoácidos, codón por codón, utilizando moléculas de tARN para añadir aminoácidos al extremo terminal de la cadena polipeptídica de la proteína que se está formando (véase la figura 1.10). Luego, cuando el ribosoma alcanza el final del mensaje, marcado por los codones de termino en el mARN, tanto el ribosoma como la proteína recien formada se desligan de mARN, dejando a éste libre en el citoplasma, listo para ser leído por otro ribosoma o para ser degradado en nucleótidos que pueden servir para otras reacciones intracelulares.

Los ribosomas operan con una eficiencia sorprendente: en 1 segundo un solo ribosoma puede añadir hasta 20 aminoácidos a la cadena polipeptídica de alguna proteína que se esté formando.

El Dogma de la Biología Molecular hasta antes de 1970.

Convendría en este punto hacer un resumen de las ideas centrales que hemos visto hasta ahora. Podemos distinguir cuatro puntos principales:

1.- La información del metabolismo celular se encuentra almacenada en el ADN nuclear a través de la secuencia de bases a lo largo de la cadena. Esta información tiene que transmitirse fielmente de la célula madre a las células hijas, para lo cual la cadena de ADN tiene que duplicarse, de modo que cada célula hija tenga la misma información genética.

2.- Esta información se copia, en el proceso de transcripción, a una molécula de ARN mensajero que lleva el mensaje genético hasta los ribosomas, fuera del núcleo.

3.- La información genética está organizada en el mARN por grupos de tres nucleótidos, llamados codones. Cada codón especifica un aminoácido.

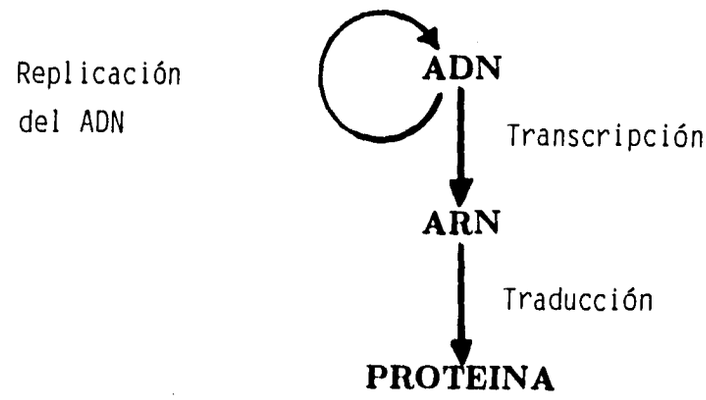
4.- En el proceso de traducción, los ribosomas leen la información contenida en el mARN utilizando moléculas adaptadoras de tARN que van añadiendo aminoácidos a la cadena polipeptídica de la proteína que se está formando, uno a la vez, dependiendo de la secuencia de codones que hay en el mARN.

Los cuatro pasos anteriores en la manipulación de la información genética fueron tan centrales en la Biología Molecular, que en su conjunto llegaron a conocerse con el nombre de "El Dogma Central de la Biología Molecular". Hasta antes de 1970 se pensaba que el camino "natural" que tenía que recorrer la información genética para poder expresarse, era el descrito por los cuatro pasos anteriores, lo cual podemos representar como

$$ADN \longrightarrow ADN \longrightarrow ARN \longrightarrow PROTEINA$$

el ADN puede copiarse en otro ADN (en la replicación celular), o puede copiarse en ARN (en la transcripción), para finalmente traducirse en proteína (en la traducción). En el esquema siguiente se muestra de forma resumida el flujo de la información genética entrañado en el Dogma Central de la Biología Molecular formulado antes de 1970 (en el capítulo siguiente hablaremos del Dogma Central en la actualidad):

Dogma Central de la Biología Molecular hasta antes de 1970



Capítulo 2

El virus del SIDA

Desde el punto de vista de las teorías físicas y químicas, los organismos biológicos no tienen el deber de existir, pero sí tienen el derecho.

Jacques Monod.

Genes.

En el capítulo anterior vimos un poco de la estructura del material genético y de su utilización: cómo está almacenada la información genética en el ADN y cómo fluye esta información desde el ADN hasta las proteínas. Vimos que la información en el ADN está almacenada a lo largo de la molécula a través de la secuencia de bases y dicha información está estructurada en codones, es decir, en triadas de bases, teniendo que cada codón especifica un aminoácido. A este nivel de estructuración de la información en el ADN la llamaremos "estructura puntual de la información". No obstante, en el ADN la información genética también está estructurada, en un nivel superior, en los *genes*. La molécula de ADN no sólo contiene la información del tipo de proteínas que utiliza la célula en su metabolismo, sino también contiene la información de los tipos de tARN's y rARN's que sirven a la célula en el proceso de traducción. Los genes son segmentos de ADN que codifican cadenas polipeptídicas (pedazos de proteínas) y ARN's.

Hasta antes de 1940 en la Biología clásica, el gen se definía como una parte del cromosoma que determinaba o especificaba un único carácter o fenotipo, por ejemplo, el color de la piel o de los ojos (fenotipo significa "aspecto externo"). Pero los trabajos realizados por George Beadle y Edward Tatum en la década de los 40's demostraron que al producir mutaciones en el ADN del hongo *neurospora crasa* (principalmente con rayos X), algunos de los mutantes obtenidos presentaban deficiencias en la producción de ciertas enzimas. Estas enzimas "defectuosas" producían un mal funcionamiento de alguna ruta metabólica determinada (una serie de reacciones químicas), lo que traía como consecuencia final que el fenotipo correspondiente a esta ruta metabólica se expresara deficientemente. Estas observaciones llevaron a la conclusión de que un gen es un segmento de ADN que

especifica o codifica una enzima, que es la hipótesis *un gen–una enzima*. Más tarde se encontró que hay genes que codifican proteínas que no son enzimas, observación que llevó a formular la hipótesis *un gen–una proteína*.

Años más tarde se descubrió que muchas proteínas constan de multiples cadenas polipeptídicas, estando cada una de estas cadenas codificada por genes diferentes. Un ejemplo típico de lo anterior estriba en la *hemoglobina* humana, proteína globular que se encuentra en las células rojas de la sangre y que es la encargada de captar oxígeno en los pulmones y transportarlo a los tejidos periféricos en donde el oxígeno se libera para efectuar la oxidación de los elementos nutritivos que producen energía . La hemoglobina está compuesta de dos tipos de cadenas polipeptídicas, las cadenas α y β, que difieren en su longitud y en su secuencia aminoacídica [†]. En la actualidad se sabe que las cadenas α y β proceden de dos grupos de genes diferentes, localizados no solamente en diferentes partes de la molécula de ADN, sino en diferentes cromosomas: en los humanos, el grupo de genes que codifica para la cadena α se encuentra en el cromosoma 16 y el que codifica para la cadena β está en el cromosoma 11. De aquí que la hipótesis "un gen–una proteína" tuvo que ser modificada a *un gen–una cadena polipeptídica.*

Sin embargo, la expresión final de los genes no siempre tiene lugar en forma de cadenas polipeptídicas. Algunos genes codifican los diferentes tipos de ARN's de transferencia mientras que otros genes especifican las diferentes clases de ARN's ribosomales. Los genes que especifican cadenas polipeptídicas y ARN's reciben el nombre de *genes estructurales*, ya que determinan la estructura de algún producto final del gen, como una enzima o un ARN estable. En el ADN existen también otros segmentos cuya función es exclusivamente reguladora. Algunos de estos segmentos no son más que señales que determinan el comienzo y el final de los genes estructurales. Podemos decir, por tanto, que el ADN contiene genes estructurales y secuencias reguladoras, en donde ahora la palabra "gen" queda determinada por la hipótesis *un gen–una cadena polipeptídica o un ARN*. A este nivel de estructuración de la información en el ADN le llamaremos "estructura global de la información genética".

Material genético no codificador.

El genoma de una célula eucariote típica contiene aproximadamente 4×10^9 nucleótidos (las células eucariotes son las que tienen núcleo, y por genoma entendemos *todo* el ADN de la célula); el tamaño de las proteínas oscila entre 100 y 2000 aminoácidos. Si suponemos que casi todo el ADN de las células eucariotes codifica para las secuencias aminoacídicas de las proteínas (sólo una pequeña parte del material genético codificador especifica las secuencias de los tARN's y rARN's) y tomando en cuenta que cada aminoácido en la proteína está codificado por tres nucleótidos, sería de esperar que en un solo organismo superior (como el ser humano) hubiera cerca de 10 millones de tipos distintos de proteínas,

[†] La estructura tridimensional de la hemoglobina fue descubierta por Max Perutz y sus colegas en Cambirdge en 1959, trabajo que le valió a Perutz el Premio Nobel de 1962.

un número probablemente inimaginable comparado con los 10 mil que se conocen en el ser humano. Una vez que se aceptó la imposibilidad de semejante número de proteínas distintas en un organismo, los biólogos comenzaron a buscar otras funciones al material genético que no fueran las de codificar para secuencias polipeptídicas. Hemos visto que en el ADN, además de los genes estructurales, existen secuencias reguladoras. Sin embargo, el tamaño de dichas secuencias (las más grandes son de 200 nucleótidos) no compensa en absoluto la gran cantidad de ADN presente en las células eucariotes.

Conforme se aprende más sobre la naturaleza específica del genoma cada vez es más difícil asignarle funciones hipotéticas a mucho del ADN. Consideremos por ejemplo el grupo de genes que codifican para la cadena β de la hemoglobina: el segmento de ADN que contiene a los 5 genes codificadores de la β−globina está compuesto por aproximadamente 75 mil nucleótidos, mientras que la cadena β de la hemoglobina sólo consta de 300 aminoácidos. Esto significa que en la codificación de la cadena β se está utilizando menos del 1% del material genético contenido en el segmento de ADN que contiene el grupo de los genes codificadores.

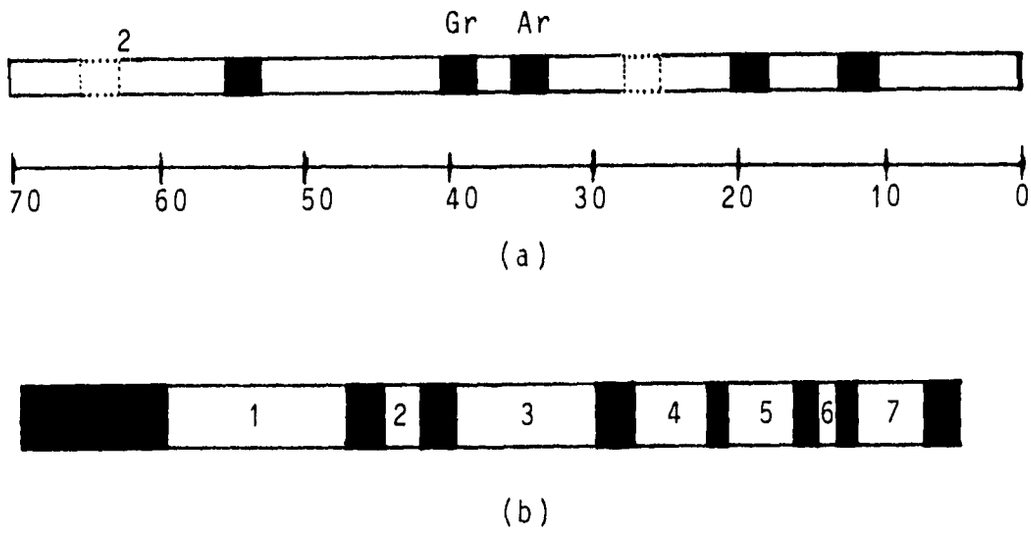


Figura 2.1

a)Mapa físico del grupo de genes de la β−globina del humano localizado en el cromosoma 11 (tomado del libro de Bioquímica de Strayer) . El mapa cubre aproximadamente 65 kilobases de ADN. Las secuencias intergénicas se representan con cajas blancas y los genes con cajas negras. Los dos genes indicados por cajas punteadas son pseudogenes. b)Estructura del gen de la ovoalbúmina, proteína de la clara de huevo de gallina. Este gen está formado por 8 exones (representados en negro) separados por 7 largos intrones (en blanco).

Efectivamente, ahora se sabe que en el ADN eucariótico existen grandes segmentos no codificadores, a los que no se les ha encontrado ninguna función, si es que la tienen. Las dos clases principales de material genético no codificador son los *intrones* y las *secuencias intergénicas* (S.I.). Las S.I. son grandes trozos de ADN que pueden contener decenas de miles de restos nucleotídicos y que se encuentran en la cadena de ADN separando los diferentes genes reguladores unos de otros. La figura 2.1(a) muestra el grupo de genes que codifican para la cadena β de la hemoglobina. Cabe mencionar aquí que en las células eucariotes más del 95% del material genético total está constituido por secuencias intergénicas. Por otro lado, los intrones son segmentos pequeños de ADN no codificador que se encuentra intercalado *dentro* de los genes. Es decir, si miramos de cerca un gen, como en la figura 2.1(b), vemos que dentro del gen existen pedazos codificadores, llamados exones, separados por pedazos no codificadores, que son los intrones. Generalmente los intrones son mucho más largos que los exones.

Cuando se descubrió la existencia de los intrones dentro de los genes, surgió una pregunta acerca del manejo de los intrones en el proceso de transcripción; claramente el mARN que sale del núcleo celular para ser leído por los ribosomas únicamente debe de llevar el mensaje genético codificador, ya que en los ribosomas se sintetiza la cadena polipeptídica especificada por el mARN, y los intrones como son secuencias no codificadoras, no podrán, por tanto, estar presentes en el mARN. De aquí que surgieran las interrogantes: ¿En la transcripción del ADN a ARN la transcriptasa se "brinca" a los intrones y sólo transcribe a los exones? ¿O se transcribe todo el gen, con exones e intrones, y después los intrones son eliminados de la cadena de ARN transcrita para dejar nada más a las secuencias codificadoras en el mARN?

Gracias a los trabajos que realizó Pierre Chambon y sus colegas de la Universidad de Estrasburgo en 1977 sobre los mecanismos involucrados en la síntesis de la proteína de la clara de huevo de gallina[5], *ovoalbúmina*, sabemos hoy que de las dos preguntas anteriores, la segunda es la que se contesta afirmativamente. En la transcripción se copia con entera fidelidad el gen completo de ADN en un largo filamento de ARN, incluidas las secuencias intrónicas. Seguidamente, ciertas moléculas enzimáticas compuestas de proteínas y de una pequeña cadena de ARN, llamadas "spliceosomas", o también *snARN* (small nuclear RNA), escinden sólo los intrones de la cadena transcrita de ARN, y reempalman con precisión el resto de la cadena, obteniendose así una molécula funcional de mARN que la célula empleará de patrón en la síntesis de alguna proteína. La función de estos spliceosomas dentro del núcleo celular fué un enigma hasta que se descubrió que sus secuencias de bases eran complementarias con las secuencias de bases existentes en los extremos de cada intrón. Por aparcamiento de las bases entre el snARN y los extremos del intrón curvado, se produce la yuxtaposición correcta de los exones, que permitirá su unión enzimática y la eliminación posterior del intrón que los separa. El snARN por lo tanto, actúa como un patrón temporal que mantiene juntos los extremos de los dos exones para que se puedan "soldar" en el punto exacto (Figura 2.2). La transformación del mARN precursor finaliza después de que todos los intrones han sido eliminados del modo indicado. Al proceso de "limpiar" al ARN transcrito quitándole los intrones se le llama *proceso de edición.*

Aunque los intrones y las S.I. son ambos material genético que no transporta información codificadora, existe una diferencia de la mayor importancia entre los dos: las S.I. nunca se transcriben, mientras que como hemos visto, los intrones sí. Otro punto importante es que en las bacterias *no* existen S.I. ni intrones. Todo el genoma bacteriano es codificador. Las S.I. y los intrones son características únicas de las células eucariotes.

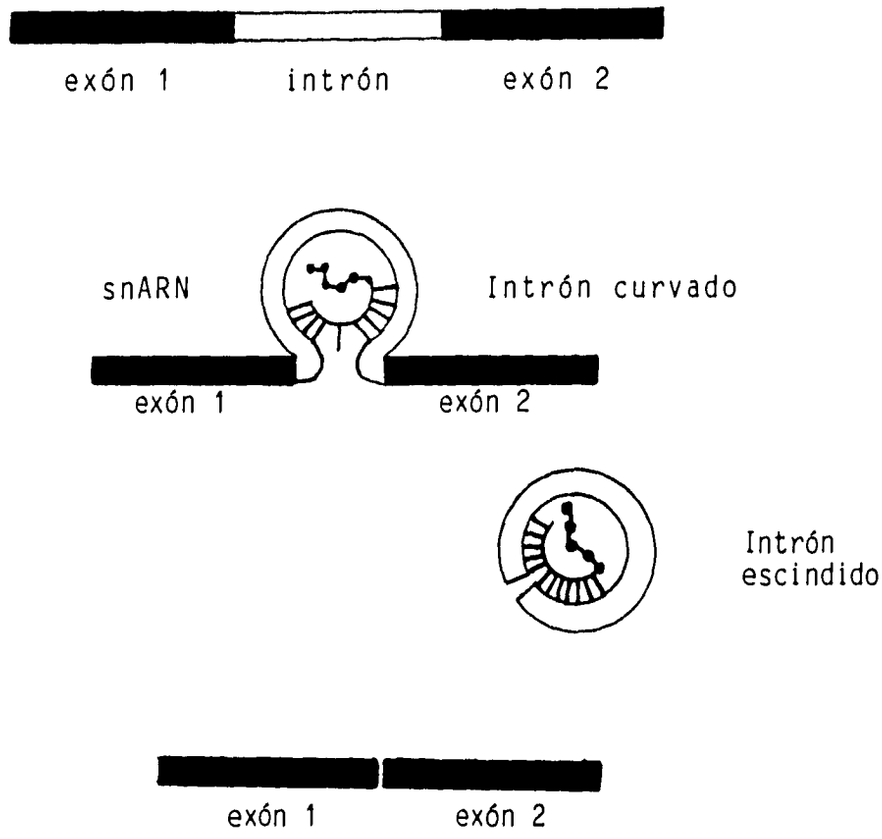


Figura 2.2

Modo de actuación del snARN en el acoplamiento de los exones en la eliminación de los intrones (proceso de edición). Los extremos del intrón aparean sus bases con un segmento del snARN. La reacción de acoplamiento va acompañada de la escisión del intrón.

Si los intrones y las S.I. son segmentos de material genético que no transportan información alguna, ¿cuál es su función dentro del genoma celular? Mucho se ha investigado al respecto, sin embargo, aún no se tienen respuestas satisfactorias. No obstante, existe evidencia que apunta en dos direcciones opuestas: por una parte, los trabajos de Thomas Cech[6] sugieren que los ARN's intrónicos tienen funciones enzimáticas y catalizadoras. Por otra parte, también hay evidencia de que en el genoma celular existen grandes cantidades de ADN que no tienen utilidad alguna y que se han acumulado en el genoma como desperdicio evolutivo. Un ejemplo de esto estriba en la salamandra, batracio urodelo que vive en sitios

oscuros y húmedos y que se alimenta principalmente de insectos. Casi genéricamente, entre más evolucionado es un organismo, más información genética requiere para controlar sus reacciones metabólicas y expresar correctamente sus características fenotípicas. Los virus, por ejemplo, tienen 10^3 veces menos información genética que las bacterias, las cuales a su vez tienen del orden de 10^3 veces menos información que las células eucariotes; los mamíferos tienen 100 veces más información que los hongos y 10 veces más información que los anfibios. Sin embargo, el genoma de la salamandra es *27 veces más grande que el de los seres humanos*. No existe ninguna razón biológica ni bioquímica del por qué la salamandra deba contener tal cantidad de ADN. No tiene funciones biológicas espectaculares (comparadas con las de los seres humanos), ni está sometida a cambios ambientales bruscos, ni nada que justifique que deba de requerir mayor información genética que nosotros. Esto hace pensar que efectivamente en el genoma de la salamandra, y si llevamos a sus últimas consecuencias el ejemplo, en el genoma de las células eucariotes, existe una gran cantidad de ADN inservible, acumulado como desperdicio a lo largo de la evolución. Cierto es que este ADN no funcional representa una carga energética para la célula en la replicación (cuando se replica, la célula madre hereda fielmente a las células hijas *todo* el genoma), pero mientras la célula haya tenido los mecanismos eficientes para replicar tanto al ADN funcional como al no funcional, no necesitó generar mecanismos que eliminaran del genoma esta "caga genética inservible". Regresaremos más tarde al análisis de las S.I.

Virus y retrovirus.

Hasta finales del siglo pasado, sólo se conocían tres características peculiares de los virus:

1.- Son partículas muy pequeñas que tienen la capacidad de atravesar los poros de los filtros utilizados para retener a las bacterias y no puede verse su estructura con el microscópio óptico.

2.- No se reproducen *in vitro*.

3.- Eran los causantes de enfermedades infecciosas como la viruela, la rabia y el sarampión.

Como puede verse, estas características son muy pobres para diferenciar nítidamente a los virus de los microorganismos celulares. Sabemos hoy que la naturaleza de los virus no es celular, y que su estructura molecular es mucho menos compleja que la de los sistemas celulares procariotes y eucariotes. Hablando genéricamente, los virus, fuera de la célula, son moléculas de ácido nucleico "encapsuladas" en una cápsula de proteínas. Más aún, no podemos siquiera decir que estén vivos, y de hecho, no lo están. Todo virus es un parásito celular que no puede reproducirse por sí mismo; necesita de la maquinaria celular para poder hacerlo. La estructura general de los virus conocida hasta antes de 1970 es la siguiente: consisten de una molécula de ADN que se encuentra, junto con algunas proteínas,

dentro de una cubierta protéica llamada *cápside*, la cual a su vez está formada por sólo unos pocos tipos diferentes de proteínas. Los mecanismos moleculares de la infección del virus a la célula huesped se conocían poco en aquella época, pero se sabía que una vez que el genoma viral entraba a célula, ésta interrumpía sus funciones normales y comenzaba a "fabricar" más y más virus. La célula no sólo hace réplicas del genoma viral, sino que también sintetiza las proteínas de la cápside y algunas otras que el virus trae consigo y que le son útilies para la infección, y las ensambla, junto con los genomas virales, para dar lugar así a la aparición de nuevas partículas víricas que pueden entonces salir de la célula para comenzar de nuevo su ciclo infeccioso.

Desde el punto de vista del Dogma Central de la Biología Molecular tratado en el capítulo anterior, no había nada nuevo acerca del flujo de la información genética en los virus y su mecanismo de infección: el ADN del virus, ya dentro de la célula, se replica para dar origen a más ADN, o se transcribe en ARN que a su vez es traducido por los ribosomas para sintetizar las proteínas virales y de la cápside. Sin embargo, en 1962, estudiando virus que producían tumores cancerígenos en las aves, Howard Temin de la Universidad de Wisconsin, propuso la existencia de virus cuyo material genético está compuesto de ARN[7]. Estos virus contienen dentro de la cápside una enzima llamada *ADN polimerasa–ARN dirigida*, o simplemente *retrotranscriptasa*, que es capaz de hacer copias de ADN a partir de la cadena molde del ARN viral. En otras palabras, la retrotranscriptasa puede transcribir la cadena de ARN viral a otra cadena de ADN. La hipótesis de la retrotranscriptasa no fué probada hasta 1970, de modo simultáneo e independiente, por el propio Temin y por David Baltimore[8], y acuñaron el nombre de *retrovirus* para estos virus cuyo material genético consta de ARN y no de ADN, como en los virus "normales".

La hipótesis de la retrotranscriptasa resultó ser de suma importancia en el entendimiento del mecanismo de infección de los retrovirus. Cuando el ARN viral entra a una célula huesped, la retrotranscriptasa lo transcribe a una molécula de ADN complementaria. Esta molécula de ADN, llamada *provirus*, tiene la capacidad de introducirse en el núcleo e insertarse en el genoma de la célula, donde queda permanentemente como un "gen" más en el ADN celular, que puede ser replicado, transcrito y traducido. La célula queda así infectada con el provirus durante todo el tiempo que dure su vida. Una vez que este ADN viral, o provirus, está insertado en el ADN nuclear de la célula, ésta puede seguir dos vías metabólicas principalmente: a) la célula transcribe y traduce su "gen viral" haciendo muchas copias del ARN del retrovirus y de sus proteínas, originando la aparición de nuevos retrovirus que pueden salir de la célula para comenzar su ciclo infeccioso en otros lugares, b) o bien, el "gen viral" permanece en el genoma celular en estado latente, sin expresarse, como si no estuviera, durante largo tiempo, que puede ir de meses a años, después del cual puede "activarse" presentándose entonces la vía a).

A los retrovirus que hacen que la célula siga la vía metabólica b) se les da el nombre de *lentivirus*, ya que después de la infección, la célula no se enferma sino hasta que haya pasado un tiempo bastante largo: estos virus actúan lentamente. En la sección siguiente describiremos detalladamente el mecanismo molecular de infección de los retrovirus, utilizando el ejemplo concreto del virus del SIDA.

El virus del SIDA.

En 1981 se conocieron los primeros casos de una nueva enfermedad llamada *síndrome de inmunodeficiencia adquirida* (SIDA). A los pacientes enfermos del SIDA se les deteriora el sistema inmune, por lo cual son víctimas de extrañas infecciones, llamadas oportunistas, las cuales son enfermedades producidas por parásitos que en condiciones normales no representan ningún peligro para el ser humano. Dos años más tarde del descubrimiento del SIDA, en 1983, Luc Montagnier y Robert Gallo identificaron las causas de la enfermedad: un virus no conocido hasta entonces, llamado *virus de la inmunodeficiencia humana* (VIH) era el causante del SIDA. Sumariamente, las características principales del VIH son las siguientes:

1.- El VIH es un retrovirus cuyo material genético consta de dos cadenas idénticas de ARN, cada una conteniendo aproximadamente 9500 residuos de nucleótidos.

2.- Pertenece a la subfamilia o grupo taxonómico *Lentivirinae*, es decir, es un lentivirus cuyo periodo de latencia es, en promedio, de 10 años.

3.- Tiene una membrana fosfolipídica externa, de tipo celular, en la cual hay incrustados 72 pares de glicoproteínas: las gp41 †, que están inmersas en la membrana, y las gp120, que se proyectan como espículas hacia el medio externo. Cada gp41 está unida a una gp120 mediante un enlace disulfuro.

4.- Debajo de la membrana viral hay una capa de proteína matricial p17, que rodea a su vez a la cápside, cuya forma es la de un cono truncado y hueco. La cápside está compuesta por sólo un tipo de proteína, la p24.

5.- Dentro de la cápside, además del ARN hay también tres tipos de proteínas enzimáticas: una integrasa, una proteasa y una retrotranscriptasa, impresindibles para la activación del virus. Además, hay otras dos proteínas, la p6 y la p7.

6.- Ataca principalmente a las células linfáticas T4, integrantes escenciales de la respuesta inmunológica, y a las células dendríticas portadoras de la proteína CD4, que se encuentran en todas las mucosas y sobre todo en los ganglios linfáticos.

En la figura 2.3 se muestra un esquema del VIH. No describiremos en su detalle la patología del virus ni las diferentes fases clínicas de la enfermedad; para ello recomendamos los exelentes artículos de Warner C. Greene y Gustavo Reyes-Terán dados respectivamente

† Las letras "gp" se refieren a "glicoproteína, y el número a su peso en kilodaltones.

en las referencias [9] y [10]. Por el momento sólo nos limitaremos a describir sumariamente cómo es que el virus infecta a la célula.

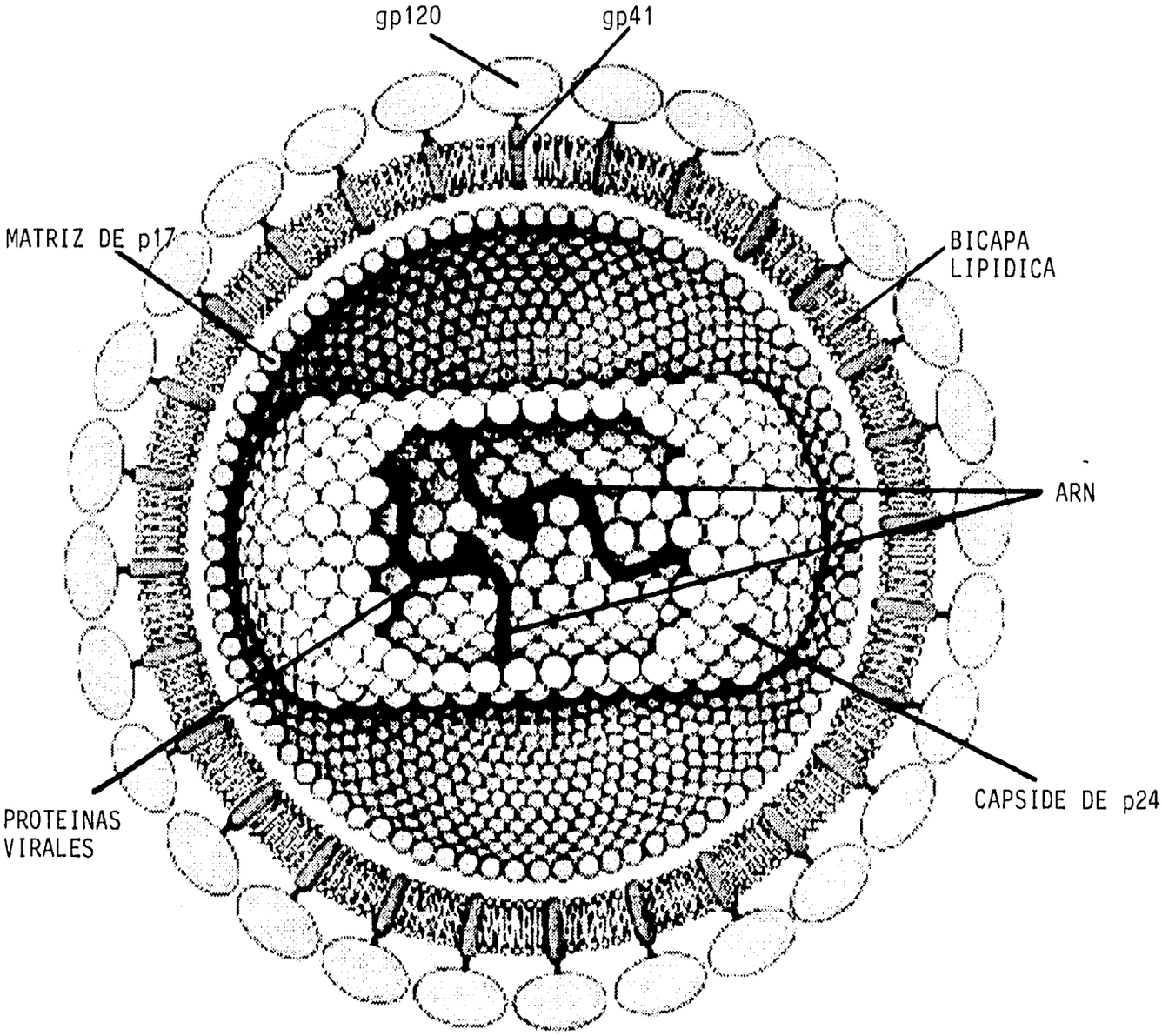


Figura 2.3

Diagrama esquemático de la estructura molecular del virus de inmunodeficiencia humana (VIH) causante del SIDA.

El virus ataca a las células portadoras de la molécula CD4 †, una proteína presente en las membranas de varios tipos de células del sistema inmune, principalmente en la superficie de los linfocitos T4 cooperadores, que son muy importantes para generar una respuesta inmunológica contra las infecciones[11]. La molécula CD4 juega el papel del receptor celular que reconoce los péptidos de proteínas ingeridas por los macrófagos o por otras células que capturan antígenos. El ciclo de la infección del VIH comienza cuando la gp120 de la envoltura viral se une a la proteína CD4 de la célula T4 cooperadora, unión que se produce de manera muy específica y con una afinidad extremadamente alta.

Después de la unión, las membranas viral y celular se fusionan –en un proceso mediado por la gp41– permitiendo la entrada de la cápside al citoplasma de la célula. Ya en el citoplasma, la cápside se abre liberando al ARN viral y a las proteínas que lo acompañan. Durante las primeras horas de la infección, el ARN viral es transcrito por la retrotranscriptasa a una cadena complementaria de ADN (provirus). El provirus se dirige entonces al núcleo celular donde se inserta en el ADN genómico de la célula huesped, quedando ahí permanentemente como un "gen" más de la célula. Este mecanismo de integración del provirus al genoma celular, llevado a cabo por la enzima viral *integrasa*, es esencial para que el virus pueda multiplicarse, y asegura que la célula quede infectada de por vida. Tal propiedad del virus hace que una persona infectada permanezca con el VIH toda su vida.

Después de la integración, el gen vírico puede persistir en un estado inactivo o latente, sin producir mensajes virales o de proteínas, lo que significa que, aparentemente, la información genética contenida en el provirus no está siendo leída y por lo tanto, la célula no replica más virus. Sin embargo, el virus está latente sólo en apariencia, y es que debemos distinguir entre las dos acepciones que tiene la palabra "latente": latencia clínica y latencia microbiológica. Desde el punto de vista clínico el virus está latente en el sentido de que el paciente infectado se siente bien, como si no estuviera enfermo. No obstante, durante el mismo periodo de latencia clínica, invariablemente todos los pacientes infectados por el VIH tienen un deterioro gradual del sistema inmune, cuya primera manifestación es el mal funcionamiento de los linfocitos T4 cooperadores y posteriormente una disminución progresiva de sus niveles circulantes. Lo anterior es una clara indicación de que no todos los virus en el paciente están completamente inactivos dentro de las células, es decir, no están en un estado de latencia microbiológica. Sin embargo, cabe mencionar que efectivamente se ha demostrado que durante algún tiempo al menos, el virus dentro de la célula permanece completamente inactivo, y son factores víricos y celulares no comprendidos aún los responsables de que el virus pierda el estado de latencia microbiológica.

De cualquier forma, tarde o temprano el "gen viral" que está insertado en el ADN celular se "activa", y en este momento, la célula huesped comienza a transcribir al provirus haciendo nuevas cadenas del ARN vírico que pueden ir al citoplasma celular, para constituir el material genético de un nuevo virus, o pueden ir a los ribosomas, donde son leídas y traducidas en las proteínas virales y de la cápside que posteriormente también formarán

† Aún no se han identificado, aunque se han propuesto, mecanismos de infección del VIH independientes de la molécula CD4

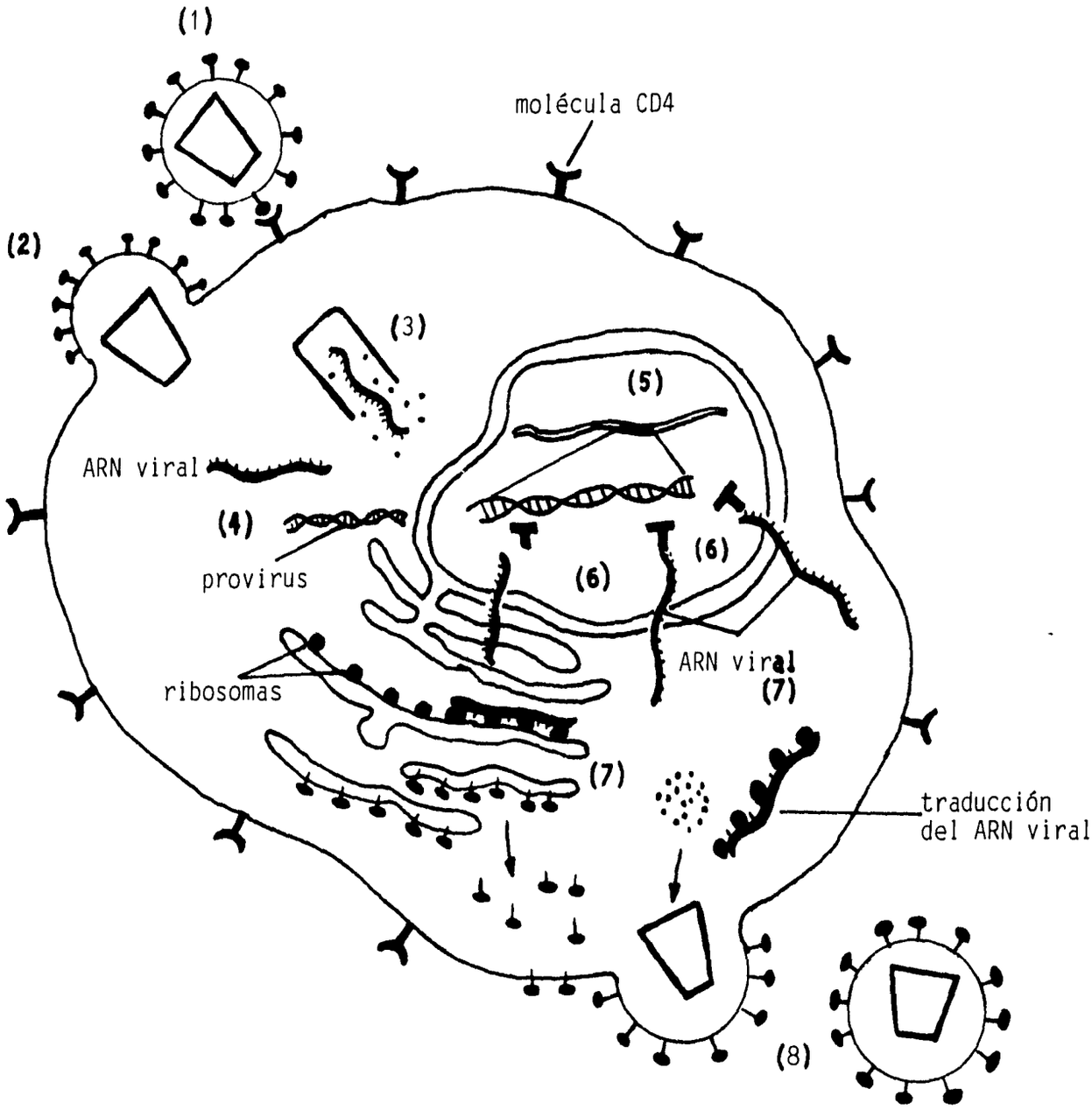


figura 2.4

Esquema del ciclo de infección del VIH. (1) El ciclo comienza cuando la gp120 se une a la proteína CD4. (2) La gp41 media el fusionamiento de las membranas celular y viral permitiendo el acceso de la cápside al citoplasma. (3) La cápside se abre liberando al ARN viral y a las proteínas que lo acompañan. (4) La retrotranscriptasa hace una copia de ADN a partir del ARN vírico (provirus). (5) El provirus entra al núcleo donde la integrasa lo inserta en el genoma celular. (6) Moléculas de la célula huesped inician la transcripción del provirus dando lugar a nuevas moléculas de ARN vírico. (7) El ARN viral abandona el núcleo para ir al citoplasma de la célula, o a los ribosomas, donde es traducido para formar las nuevas proteínas virales. (8) La proteasa es la encargada de ensamblar a los ARN's y proteínas virales, dando lugar así a la gemación de un nuevo virus (8).

parte de un nuevo virus. En un proceso mediado por la enzima viral *proteasa*, en el citoplasma de la célula se lleva a cabo el ensamblaje del núcleo vírico, es decir, de la cápside junto con el ARN del retrovirus y las proteínas asociadas a él. Una vez ensamblado, el centro del nuevo retrovirus se dirige a la superficie y sale a través de la menbrana celular, de la que adquiere su propia membrana y se completa con las proteínas principales externas: la gp41 y la gp120. En este momento un nuevo VIH queda completamente formado, y es capaz de ir a infectar a otra célula comenzando de nuevo su ciclo de replicación.

Ya que hemos visto con algún detalle el mecanismo de infección del VIH, estamos en disposición de describir la estructura génica del genoma viral (figura 2.5). El genoma del VIH contiene a tres genes estructurales que son comunes en todos los virus pertenecientes a la familia *Retroviridae* (retrovirus): los genes *env, gag* y *pol*. El gen *env* codifica para las glicoproteínas gp120 y gp41 de la membrana viral que median la unión a la CD4 y la fusión con la membrana celular, respectivamente, mientras que el gen *gag* codifica para la proteína p24 con la que está hecha la cápside y para la p17 de la matriz. Por otro lado, el gen *pol* contiene la información de las proteínas enzimáticas retrotranscriptasa, proteasa, integrasa y ribonucleasa. Además, el genoma viral contiene tres genes con actividad reguladora, *tat, rev* y *nef*, y tres con funciones accesorias, *vif, vpu* y *vpr*. El gen *tat* codifica para una proteína que acelera en un factor de 1000 la velocidad de retrotranscripción de la ARN polimerasa–ADN dirigida, mientras que el gen *rev* es regulador de la expresión de los genes estructurales. Por otro lado, la función del gen *nef* es aún incierta, aunque hay pruebas que indican que la proteína codificada por este gen modifica a la célula para la fabricación de nuevos VIH.

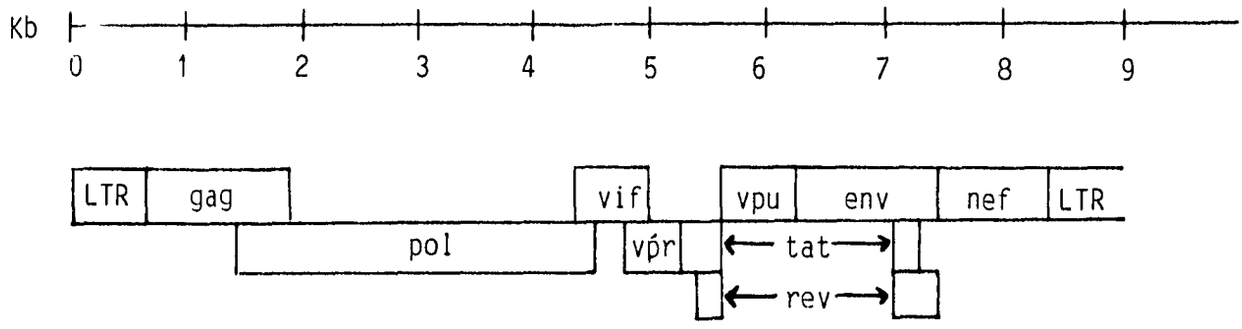
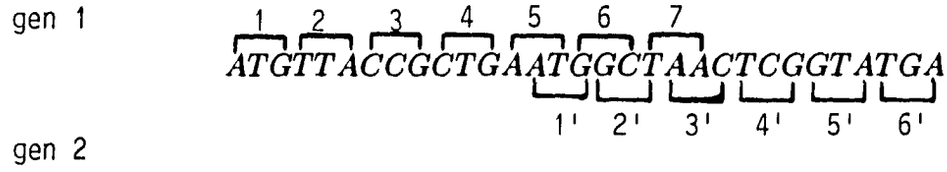


Figura 2.5

Los genes del VIH están indicados por las posiciones de las diferentes cajas a lo largo del ADN del provirus. Los genes que se traslapan utilizan la misma region del ADN, pero son leídos de forma distinta por los ribosomas de la célula huesped.

El papel que juegan las tres proteínas codificadas por los genes *vpr, vif* y *vpu* en la capacidad de infección del virus no ha quedado claro aún; tal vez los dos últimos tengan alguna función en el ensamblaje de los nuevos virus, mientras que el gen *vpr* sea un activador transcripcional débil, no tan potente como el *tat*.

La figura 2.5 muestra esquemáticamente como están acomodados estos genes a lo largo del genoma viral. Notemos de la figura que estos genes están traslapados, es decir, un gen comienza antes de que se termine el otro. La región traslapada de ADN en dos genes distintos es utilizada por ambos genes, sin embargo, es leída en marcos de lectura distintos por la maquinaria de síntesis protéica de la célula huesped. Recordemos que todo gen comienza con el codón de inicio ATG, y termina con alguno de los codones de término TAA, TGA o TAG. Por ejemplo, en la secuencia



tenemos dos genes traslapados que son leídos en diferentes marcos de lectura. El codón 7 es un codón de término en el marco de lectura del gen 1, pero no lo es en el marco de lectura del gen 2. Por otro lado, el codón 1' es un codón de inicio para el gen 2, mientras que no tiene significado en el marco de lectura del gen 1. Se dice que ambos marcos de lectura están *defasados*. De esta manera se produce el traslape de genes en el VIH.

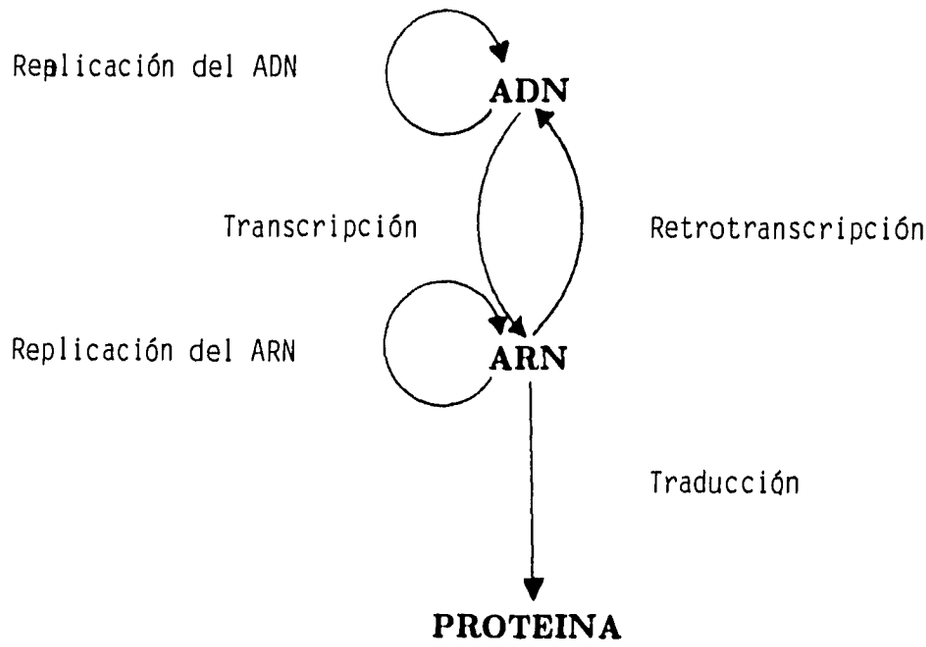
El Dogma Central de la Biología Molecular en la actualidad.

En el capítulo anterior vimos cuál era la manera "natural" en que fluía la información genética que se conocía antes de 1970. Sin embargo, en esa década los trabajos de Temin y de Baltimore demostraron que en los retrovirus la información genética también podía fluir "al revés", de ARN a ADN. Más aún, en la misma década se demuestra que existen virus, llamados ahora *reovirus*, cuyo material genético está compuesto de ARN, y que portan consigo una enzima ARN polimerasa–ARN dirigida que es capaz de transcribir moléculas de ARN copiándolas a otras moléculas complementarias también de ARN. De modo que ahora la información genética también puede fluir de ARN a ARN.

Estos trabajos, iniciados por Temin en 1962, hicieron necesaria la reformulación del Dogma Central de la Biología Molecular para que éste tomara en cuenta a los mecanismos

de replicación de los retrovirus y los reovirus. Así, la extensión posterior del Dogma Central de la Biología Molecular sobre el flujo de la información genética puede resumirse como se muestra en el siguiente esquema:

Dogma Central de la Biología Molecular Actual



Capítulo 3

El ADN y periodo 3

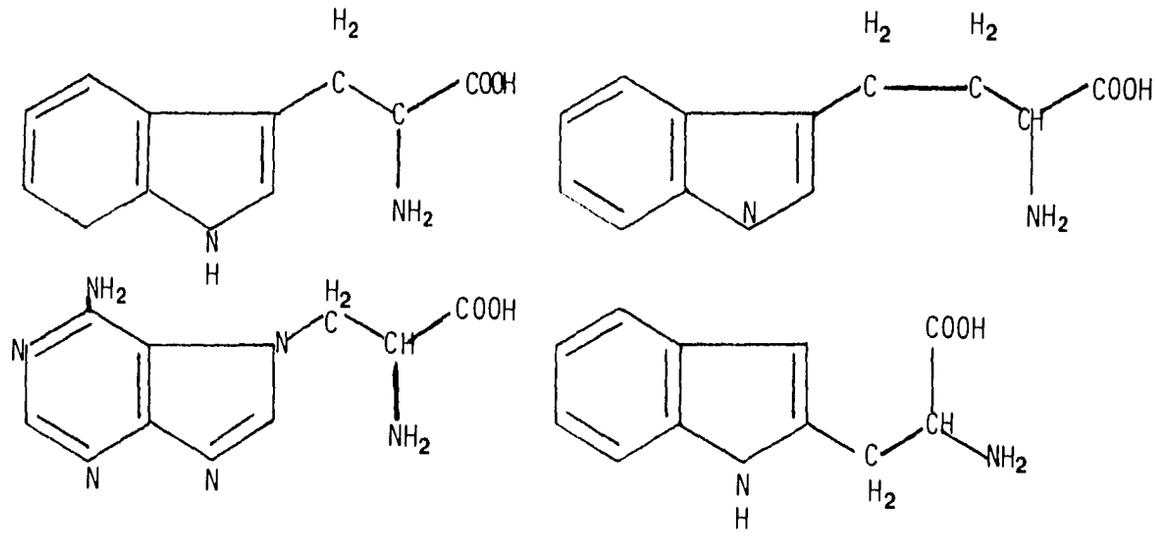
Esta creencia, de que la evolución "darwiniana" está hecha al azar, no es sólo falsa. Es exactamente lo opuesto a la verdad. El azar es sólo un pequeño ingrediente de la receta darwiniana, pero el ingrediente más importante es la selección cumulativa, cuya quintaesencia es, precisamente, que no está hecha al azar.
Richard Dawkins.

La degeneración del código

La naturaleza del código genético quedó completamente dilucidada en 1966. Tal vez la característica más importante del código genético es la de que sea un código degenerado, es decir, que diferentes codones especifican el mismo aminoácido. Recordemos que existen 61 codones que especifican solamente 20 aminoácidos (3 de los 64 codones existentes son codones de termino que no especifican ningún aminoácido). Tal degeneración del código puede tener un posible significado biológico: pudiera ser que la degeneración reduce al mínimo los efectos deletéreos de las mutaciones. Resulta que en los procesos de replicación o de transcripción, las enzimas encargadas de llevar a cabo dichos procesos pueden cometer cierto tipo de "errores", llamados mutaciones deletéreas, que consisten en que tales enzimas intercambian una base por otra. Por ejemplo, en la transcripción la transcriptasa en lugar de añadir una C a la cadena de ARN en formación, se "equivoca" y añade una U. Si el código genético no fuera degenerado, 20 codones especificarían univocamente 20 aminoácidos, y los otros 44 codones restantes servirían como codones de término, por ejemplo. En este caso hipotético, una mutación originaría o bien el cambio de un aminoácido en la cadena polipeptídica codificada, o la terminación prematura de la cadena, obteniendose así una proteína no funcional. Cabe mencionar que la *anemia falciforme*, una enfermedad de los globulos rojos de la sangre que puede resultar mortal, tiene como causa una mutación deletérea que origina el cambio de uno sólo de los casi 600 aminoácidos de la hemoglobina. De modo que la degeneración del código genético puede deberse a una "estrategia" de la evolución que reduce la probabilidad de que al cambiar una "letra" por otra en un codón, cambie el aminoácido especificado por dicho codón.

Una vez que se dilucidó completamente la naturaleza del código genético, surgió una pregunta de la mayor importancia: ¿Por qué los codones están constituidos por tres bases? Es decir, ¿Por qué la información contenida en el ADN se lee de tres en tres, y no de dos en dos, por ejemplo, o de cuatro en cuatro? La respuesta que se le dio a la pregunta anterior fue la siguiente: si la información en el ADN se leyera de dos en dos, sólo habría $4 \times 4 = 16$ diadas de bases, que no son suficientes para especificar a los 20 aminoácidos. Por otro lado, si la información se leyera de cuatro en cuatro, habría un total de $4 \times 4 \times 4 \times 4 = 256$ cuaternas de bases, y 256 cuaternas son demasiadas para especificar a solamente 20 aminoácidos, lo que conduciría a una degeneración excesiva del código genético. De modo que la manera óptima de leer la información contenida en la cadena de ADN es de tres bases en tres bases. El número 3 representa, entonces, el cumplimiento con los compromisos de "suficiencia" y de "degeneración no excesiva" en el código genético.

No obstante, esta respuesta a la pregunta del por qué el "número mágico 3" en la cadena de ADN es, hasta cierto punto, ingenua y de poca profundidad, ya que asume tácitamente que en el proceso evolutivo surgieron primero las proteínas, formadas por los 20 aminoácidos, y después apareció el ADN, cuya información contenida tuvo que estructurarse y amoldarse a los 20 aminoácidos ya existentes, lo cual no parece estar de acuerdo con el flujo de la información genética observado en los organismos vivos. Más aún, el contestar de este modo al por qué la presencia del número 3 en el ADN nos conduce directamente al surgimiento de otro número mágico: el 20. En otras palabras, ¿Por qué entonces las proteínas están hechas por 20 y sólo 20 aminoácidos? En efecto, de los cuatro aminoácidos siguientes



solamente el primero (triptofano) es sintetizado por las células, aunque los cuatro tienen propiedades químicas muy similares.

```
   1 ATGAAAGTGA AGGGGATCAG GAAGAATTAT CAGCACTTGT GGAAATGGGG CATCATGCTC
  61 CTTGGGATGT TGATGATCTG TAGTGCTGTA GAAAATTTGT GGGTCACAGT TTATTATGGG
 121 GTACCTGTGT GGAAAGAAGC AACCACCACT CTATTTTGTG CATCAGATGC TAAAGCATAT
 181 GATACAGAGG TACATAATGT TTGGGCCACA CATGCCTGTG TACCCACAGA CCCCAACCCA
 241 CAAGAAGTAG TATTGGAAAA TGTGACAGAA AATTTTAACA TGTGGAAAAA TAACATGGTA
 301 GAACAGATGC ATGAGGATAT AATCAGTTTA TGGGATCAAA GCCTAAAGCC ATGTGTAAAA
 361 TTAACCCCAC TCTGTGTTAC TTTAAATTGC ACTGATTTGA GGAATGTTAC TAATATCAAT
 421 AATAGTAGTG AGGGAATGAG AGGAGAAATA AAAAACTGCT CTTTCAATAT CACCACAAGC
 481 ATAAGAGATA AGGTGAAGAA AGACTATGCA CTTTTTTATA GACTTGATGT AGTACCAATA
 541 GATAATGATA ATACTAGCTA TAGGTTGATA AATTGTAATA CCTCAACCAT TACACAGGCC
 601 TGTCCAAAGG TATCCTTTGA GCCAATTCCC ATACATTATT GTACCCCGGC TGGTTTTGCG
 661 ATTCTAAAGT GTAAAGACAA GAAGTTCAAT GGAACAGGGC CATGTAAAAA TGTCAGCACA
 721 GTACAATGTA CACATGGAAT TAGGCCAGTA GTGTCAACTC AACTGCTGTT AAATGGCAGT
 781 CTAGCAGAAG AAGAGGTAGT AATTAGATCT AGTAATTTCA CAGACAATGC AAAAAACATA
 841 ATAGTACAGT TGAAAGAATC TGTAGAAATT AATTGTACAA GACCCAACAA CAATACAAGG
 901 AAAAGTATAC ATATAGGACC AGGAAGAGCA TTTTATACAA CAGGAGAAAT AATAGGAGAT
 961 ATAAGACAAG CACATTGCAA CATTAGTAGA ACAAAATGGA ATAACACTTT AAATCAAATA
1021 GCTACAAAAT TAAAAGAACA ATTTGGGAAT AATAAAACAA TAGTCTTTAA TCAATCCTCA
1081 GGAGGGGACC CAGAAATTGT AATGCACAGT TTTAATTGTG GAGGGGAATT TTTCTACTGT
1141 AATTCAACAC AACTGTTTAA TAGTACTTGG AATTTTAATG GTACTTGGAA TTTAACACAA
1201 TCGAATGGTA CTGAAGGAAA TGACACTATC ACACTCCCAT GTAGAATAAA ACAAATTATA
1261 AATATGTGGC AGGAAGTAGG AAAAGCAATG TATGCCCCTC CCATCAGAGG ACAAATTAGA
1321 TGCTCATCAA ATATTACAGG GCTAATATTA ACAAGAGATG GTGGAACTAA CAGTAGTGGG
1381 TCCGAGATCT TCAGACCTGG GGGAGGAGAT ATGAGGGACA ATTGGAGAAG TGAATTATAT
1441 AAATATAAAG TAGTAAAAAT TGAACCATTA GGAGTAGCAC CCACCAAGGC AAAAAGAAGA
1501 GTGGTGCAGA GAGAAAAAAG AGCAGTGGGA ACGATAGGAG CTATGTTCCT TGGGTTCTTG
1561 GGAGCAGCAG GAAGCACTAT GGGCGCAGCG TCAATAACGC TGACGGTACA GGCCAGACTA
1621 TTATTGTCTG GTATAGTGCA ACAGCAGAAC AATTTGCTGA GGGCTATTGA GGCGCAACAG
1681 CATCTGTTGC AACTCACAGT CTGGGGCATC AAGCAGCTCC AGGCAAGAGT CCTGGCTCTG
1741 GAAAGATACC TAAGGGATCA ACAGCTCCTA GGGATTTGGG GTTGCTCTGG AAAACTCATC
1801 TGCACCACTG CTGTGCCTTG GAATGCTAGT TGGAGTAATA AAACTCTGGA TATGATTTGG
1861 GATAACATGA CCTGGATGGA GTGGGAAAGA GAAATCGAAA ATTACACAGG CTTAATATAC
1921 ACCTTAATTG AAGAATCGCA GAACCAACAA GAAAAGAATG AACAAGACTT ATTAGCATTA
1981 GATAAGTGGG CAAGTTTGTG GAATTGGTTT GACATATCAA ATTGGCTGTG GTATATAAAA
2041 ATCTTCATAA TGATAGTAGG AGGCTTGATA GGTTTAAGAA TAGTTTTTAC TGTACTTTCT
2101 ATAGTAAATA GAGTTAGGCA GGGATACTCA CCATTGTCAT TTCAGACCCA CCTCCCAGCC
2161 CCGAGGGGAC CCGACAGGCC CGAAGGAATC GAAGAAGAAG GTGGAGACAG AGACAGAGAC
2221 AGATCCGTGC GATTAGTGGA TGGATTCTTA GCACTTTTCT GGGACGACCT GCGGAGCCTG
2281 TGCCTCTTCA GCTACCACCG CTTGAGAGAC TTACTCTTGA TTGTAGCGAG GATTGTGGAA
2341 CTTCTGGGAC GCAGGGGGTG GGAAGTCCTC AAGTATTGGT GGAATCTCCT GCAGTATTGG
2401 AGTCAGGAAC TAAGGAATAG TGCTGTTAGC TTGCTTAATG CCACAGCTAT AGCAGTAGCT
2461 GAGGGGACAG ATAGGGTTAT AGAAGTAGTA CAAAGAATTT ATAGGGCTAT TCTCCACATA
2521 CCTACAAGAA TAAGACAGGG CTTGGAAAGG CTTTTGCTAT AA
```

Figura 3.1

Secuencia genética del gen env de un virus de SIDA tipo 1. La longitud de la secuencia es de 2562 bases (tomada del Gene Bank 94).

En este trabajo no intentaremos contestar contundentemente a la pregunta de por qué el 3 en el ADN. No obstante, en el capítulo presente expondremos un criterio de plausibilidad que conlleva a la conclusión de que la presencia del número 3 en el ADN está lejos de ser fortuita. Para esto consideraremos a la molécula de ADN (y a la de ARN también) bajo el punto de vista únicamente de la información que contiene. En otras palabras, de ahora en adelante olvidaremos la estructura química y molécular de la cadena de ADN presentada en el capítulo 1, y la consideraremos como una simple secuencia de letras (A, T, C, G), tal y como se muestra en la figura 3.1.

Representación binaria del ADN.

En el capítulo 1 vimos que las bases A y T se enlazan complementariamente utilizando dos puentes de hidrógeno, mientras que las bases C y G se enlazan a través de tres puentes de hidrógeno. Por esta razón, al enlace complementario A–T se le llama *enlace débil*, mientras que al enlace complementario C–G se le llama *enlace fuerte*. En base a esto, nosotros además adoptaremos la terminología de llamar a las bases A y T "bases débiles" y a las bases C y G las llamaremos "bases fuertes". Esta clasificación de las bases es binaria ya que estamos agrupando a las cuatro bases A, T, C y G en sólo dos posibilidades: o son débiles o son fuertes. Lo anterior nos conduce a una representación binaria de la cadena de ADN que, lejos de ser arbitraria, tiene un profundo significado físico: en la cadena de ADN sustituiremos a las bases débiles A y T por ceros y a las bases fuertes C y G por unos:

$$\left\{\begin{matrix} A \\ T \end{matrix}\right\} \rightarrow 0$$

$$\left\{\begin{matrix} C \\ G \end{matrix}\right\} \rightarrow 1$$

Por ejemplo, bajo esta representación, la secuencia

$$ATTCTGGCCATACTGCAT$$

se convierte en

$$0\,0\,0\,1\,0\,1\,1\,1\,1\,0\,0\,0\,1\,0\,1\,1\,0\,0.$$

Con miras a encontrar una regularidad en las secuencias genéticas, y en particular en aquellas pertenecientes al VIH, tomamos la Transformada de Fourier Discreta (TFD) a dichas secuencias en la representación binaria. Recordemos que si $x_1, x_2, \cdots, x_n$ es una secuencia discreta de datos, la TFD de éstos es la operación que crea una correspondiente

secuencia de datos $\hat{x}_k$ tales que

$$\hat{x}_k = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} x_j \exp\left(-i\frac{2\pi jk}{n}\right) \qquad (1)$$
$$k = 1,\ 2,\ \cdots,\ n$$
$$i = \sqrt{-1}$$

Podemos considerar la transformación definida en (1) como un tipo de rotación que mapea el vector $(x_1, x_2, \ldots, x_n)$ al vector $(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$. No es una rotación en el sentido usual, debido a que la longitud euclidiana $\sum_j x_j^2$ no se conserva. Lo que se conserva bajo la transformación (1) es la longitud *hermitiana*†, tal y como se expresa en la ecuación de Parseval-Plancherel:

$$\sum_{j=1}^{n} |x_j|^2 = \sum_{j=1}^{n} |\hat{x}_j|^2$$

En el caso de las secuencias genéticas en la representación binaria, tomamos los x_j como los dígitos (0 o 1) que aparecen a lo largo de la secuencia, siendo x_1 el valor del primer dígito, x_2 el del segundo, y así sucesivamente.

La TFD nos revela las regularidades y estructuras que existen en la secuencia discreta de datos $x_1,\ x_2,\ \ldots,\ x_n$, a través del *espectro de potencias*, que no es más que la gráfica de $|\hat{x}_k|^2$ contra k. El espectro de potencias es simétrico respecto al punto $k = n/2$, ya que de la definición (1) se ve que

$$\hat{x}_k = \hat{x}^*_{n-k}$$

y por esta razón, en nuestras gráficas sólo mostraremos la mitad del espectro de potencias, es decir, la gráfica de $|\hat{x}_k|^2$ contra k para $k = 1, 2, \ldots, n/2$, teniendo en cuenta que la otra mitad es simétrica con la anterior.

En la figura 3.2 se muestra el espectro de potencias correspondiente a la representación binaria de la secuencia genética de todo el genoma de un virus de inmunodeficiencia humana tipo 1 (VIH−1)‡. La longitud de esta secuencia era de 9229 bases, de modo que el pico que aparece tan marcadamente en el espectro de potencias de la figura 3.2, en la posición $k = 3050$, refleja una periodicidad en la cadena de

$$\frac{9229}{3050} = 3.026$$

---

† La longitud hermitiana es la cantidad $\sum xx^* = \sum |x|^2$, donde x^* denota el complejo conjugado de x

‡ Existen dos tipos de virus de SIDA, llamados VIH-1 y VIH-2. El VIH-2 es el virus común causante del SIDA en los paises africanos, mientras que el VIH-1 es el causante del SIDA en los paises americanos. La estructura molecular de ambos virus es la misma que la presentada en la figura 2.3, y su estructura génica es como la mostrada en la figura 2.5. En lo que se diferencían es en que el VIH-2 causa la muerte más rápido de los pacientes infectados que el VIH-1.

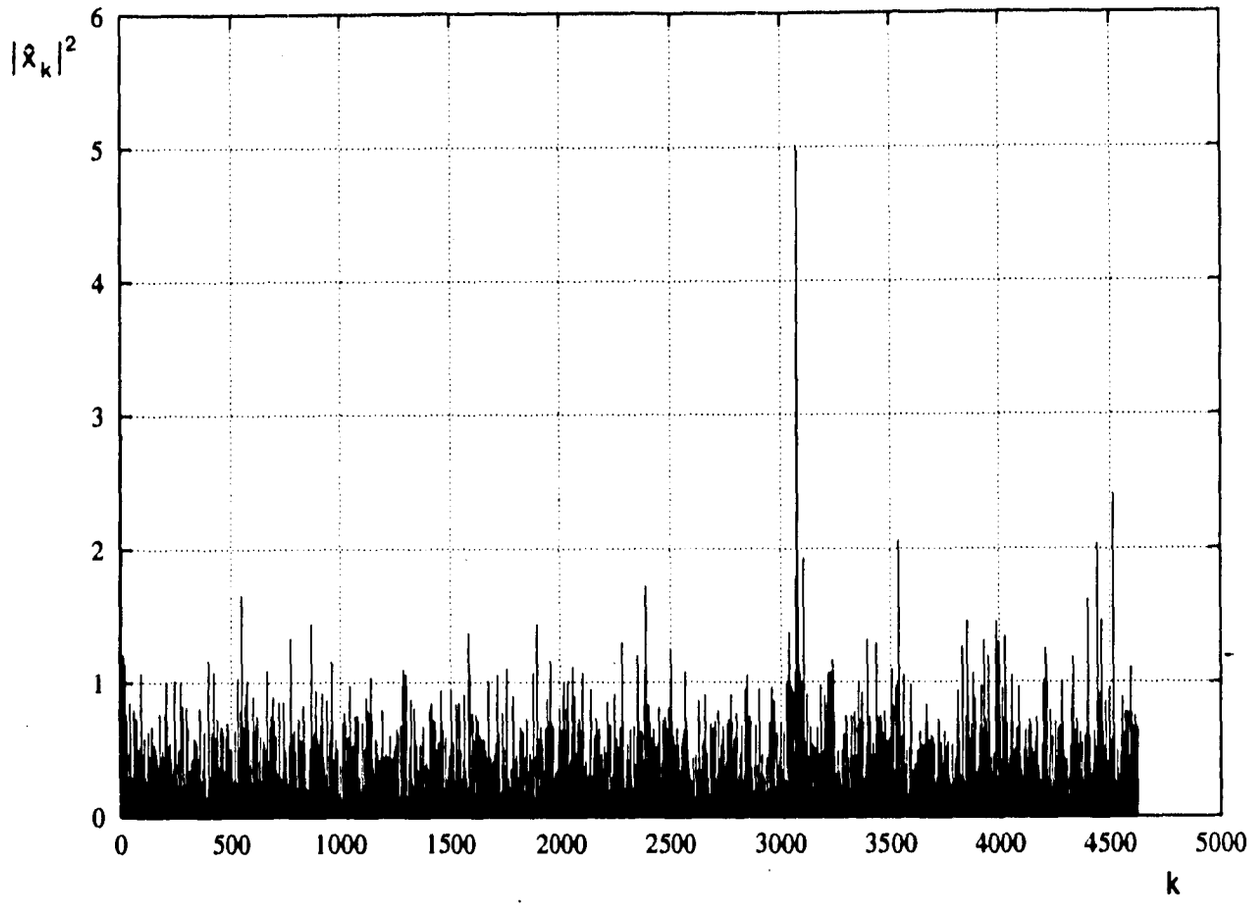


figura 3.2

Espectro de potencias correspondiente a la representación binaria de la secuencia genética del genoma de un virus de SIDA tipo 1. Sólo se muestra la mitad del espectro, ya que la otra mitad es simétrica.

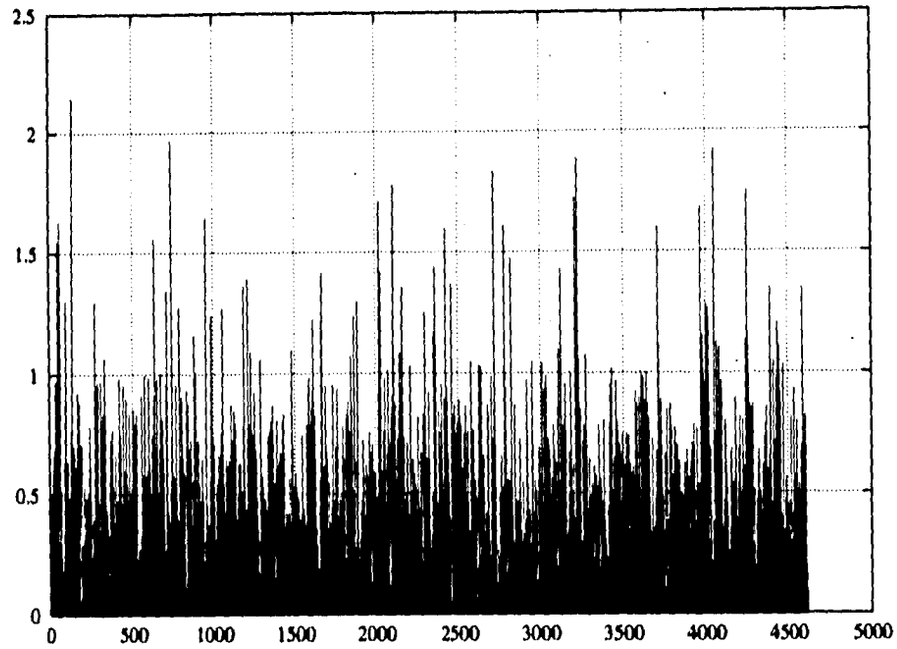
Para tener parámetros de comparación generamos con la computadora "secuencias genéticas" aleatorias (mejor dicho "pseudoaleatorias", ya que el generador de números aleatorios que usamos no genera números aleatorios cien por ciento "puros"), de la misma longitud que la secuencia genética verdadera (9229 bases). También construimos una "secuencia genética", de la misma longitud, completamente periódica, siguiendo el patrón

$$AATTCCGGAATTCCGGAATTCCGG\ldots$$

lo que conduce, bajo nuestra representación, a la secuencia binaria

$$0\,0\,0\,0\,1\,1\,1\,1\,0\,0\,0\,0\,1\,1\,1\,1\,0\,0\,0\,0\,1\,1\,1\,1\ldots$$

(a)



(b)

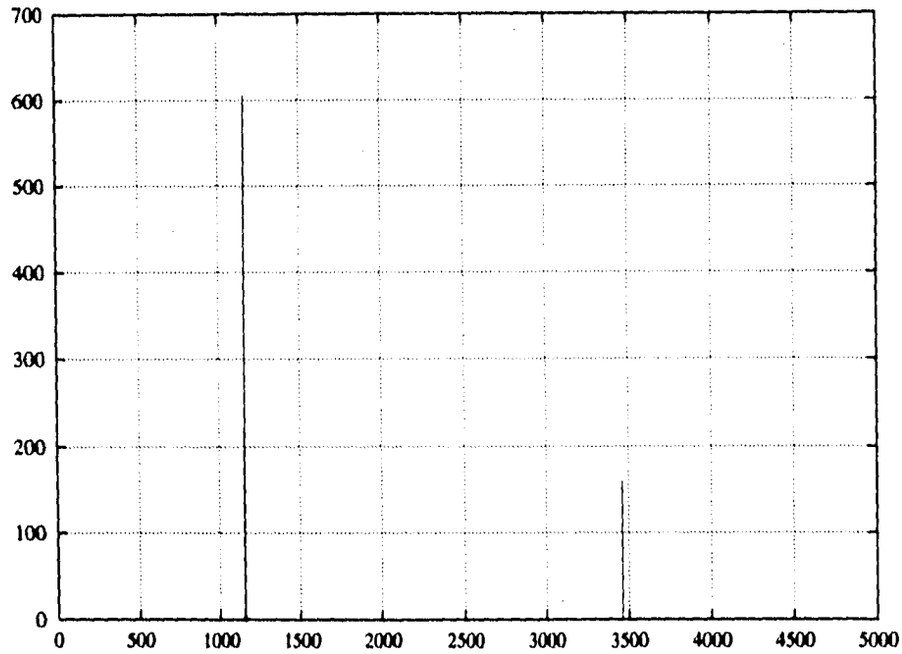


figura 3.3

(a)Espectro de potencias correspondiente a una cadena de ADN aleatoria. (b) Espectro de potencias de una cadena de ADN periódica con periodo 8.

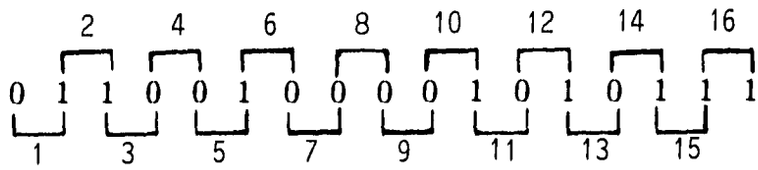
En las figuras 3.3(a) y 3.3(b) se muestran los espectros de potencias de las cadenas aleatoria y periódica, respectivamente. Comparando estas gráficas con la dada en la figura 3.2, vemos que la secuencia genética del VIH-1, lejos de ser aleatoria, es más bien periódica, con periodo 3; periodicidad que se manifiesta a través de la representación binaria que hemos escogido para el ADN.

Hasta el momento sólo hemos puesto de manifiesto la presencia del periodo 3 en las secuencias genéticas, pero no la hemos justificado. Es decir, con la TFD únicamente hemos "exhibido" al periodo 3 explícitamente en las secuencias genéticas, pero vemos, no obstante, que en las cadenas aleatorias esta periodicidad no está presente, al menos no con la representación binaria que hemos escogido para el ADN. En la sección siguiente mostraremos otro método que muestra que también en las cadenas aleatorias se tiene cierto tipo de periodo 3.

Mínimos de energía en cadenas binarias.

Continuemos con el análisis de las cadenas binarias de ceros y unos, pero ya no pensando en ellas como la representación energética fuerte-débil del ADN vista en el apartado anterior. Pensemos ahora en estas cadenas binarias simplemente como lo que son: cadenas binarias. O mejor aún, para dramatizar más las cosas, imaginemos el caso hipotético en el que tenemos un "ADN" compuesto de sólo dos "bases", 0 y 1, e imaginemos que nuestras cadenas binarias representan las "secuencias genéticas" de este "ADN binario", por llamarlo de alguna forma.

Supongamos además que existe una energía de interacción entre una pareja de bases consecutivas en este "ADN binario" tomadas a lo largo de la cadena, con la pareja complementaria que se encuentra en la segunda cadena (recordemos que el ADN consta de dos cadenas, las cuales son complementarias una con respecto a la otra). Si las bases que forman nuestro ADN binario siguen un princípio de complementariedad, es decir, que el 0 sólo se ligue con el 1 (tal como la A sólo se liga con la T), podemos hablar, por ejemplo, de la energía E_{01} necesaria para romper los enlaces que ligan a la pareja "01" con la pareja complementaria "10", o la energía E_{11} necesaria para separar a la pareja de bases "11" de la pareja complementaria "00". Aunque estemos presentando el problema como si fueran cadenas unidimensionales, debemos recordar que estas energías de interacción se dan entre parejas de bases que son complementarias y que se encuentran en dos cadenas diferentes. A este tipo de interacciones las llamaremos *interacciones de diadas.* En la cadena binaria siguiente mostramos un ejemplo de cómo tomaremos las diadas, o parejas de bases, a lo largo de la cadena (recordando que esta cadena está ligada a otra, que no mostramos, pero que es complementaria a la que sí se muestra), y hemos numerado las diadas dependiendo de su posición, contando desde el primer dígito, en dicha cadena:



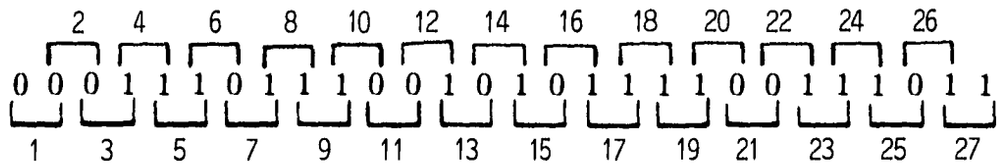
(2)

En el caso binario podemos formar 4 diadas, a saber, 00, 01, 10 y 11, y por lo tanto tendremos cuatro energías de interacción correspondientes: E_{00}, E_{01}, E_{10}, y E_{11}. Supongamos que los valores numéricos de estas energías satisfacen las relaciones

$$E_{00} = E_{11} > E_{01} > E_{10} \tag{3}$$

(más adelante quedará claro el por qué de estas relaciones). Para concretar, asignemos los siguientes valores numéricos a las energías, valores que, aunque son arbitrarios, satisfacen la relación (3):

$$\begin{aligned} E_{00} = E_{11} &= 9 \\ E_{01} &= 5 \\ E_{10} &= 1 \end{aligned} \tag{4}$$



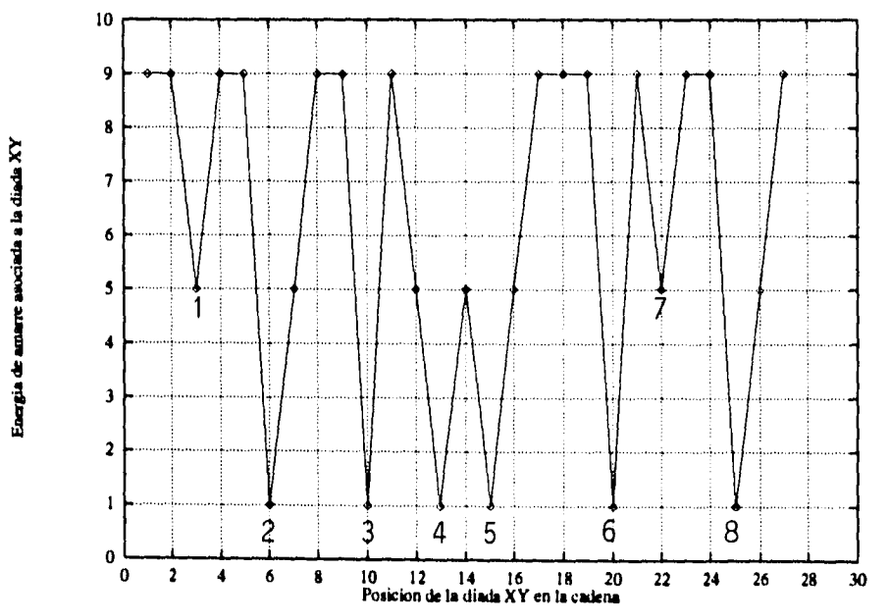


Figura 3.4

En esta figura se muestra una secuencia binaria aleatoria de 28 dígitos de longitud. Abajo, la gráfica de energías correspondiente a esta secuencia. Se han numerado los mínimos de energía para referencia en el texto

Si ahora llamamos n al número que representa la posición en la cadena de la diada XY contando desde el comienzo de la cadena, como se muestra en el ejemplo (2), podemos hacer la gráfica E_{XY} vs n, es decir, la gráfica de la energía de interacción asociada a la diada XY, contra el número de posición de dicha diada a lo largo de la cadena. En la figura 3.4 se muestra una cadena binaria y la gráfica de energías correspondiente a esta cadena. En dicha gráfica tenemos 8 mínimos numerados, 1, 2,...,8, y lo que nos interesa es la *distancia promedio* que hay entre dos mínimos consecutivos. Por ejemplo, el mínimo 1 se encuentra en la posición $n_1 = 3$, mientras que el mínimo 2 se encuentra en $n_2 = 6$. Por lo tanto, la distancia d_{12} entre los mínimos 1 y 2 es

$$d_{12} = n_2 - n_1 = 3.$$

Por otro lado, el mínimo 3 está en la posición $n_3 = 10$, de modo que la distancia d_{23} entre los mínimos 2 y 3 es

$$d_{23} = n_3 - n_2 = 4.$$

Análogamente calculamos las distancias d_{34}, d_{45},..., d_{78}, y la distancia promedio entre mínimos de energía sucesivos $\bar{d}$, la definimos como el promedio aritmético de todas estas distancias:

$$\bar{d} = \frac{d_{12} + d_{23} + \cdots + d_{m-1,m}}{m-1} \tag{5}$$

siendo m el número total de mínimos que hay a lo largo de toda la cadena. Es fácil ver que en el caso de la figura 3.4 la distancia promedio entre mínimos es

$$\bar{d} = 3.14$$

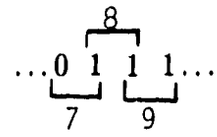
Cabe mencionar que la cadena binaria de la figura 3.4 fue generada al azar. En cualquier caso, esta cadena es muy pequeña como para obtener de ella resultados significativos, así que construímos varias cadenas binarias aleatorias, cada una de 100000 dígitos de longitud, y utilizando los valores de energías asignados en (4), calculamos las distancias promedio entre mínimos de energía consecutivos para cada una, obteniendo que para todas ellas la distancia promedio fue

$$\bar{d} = 3.20 \pm 0.02 \tag{6}$$

Este resultado podemos probarlo haciendo cálculos probabilísticos de la manera siguiente. Definamos $E_{XY}(n)$ como la energía asociada a la diada XY que se encuentra en la posición n en la cadena. Para saber si $E_{XY}(n)$ se encuentra en un mínimo, es necesario comparar su valor con el de $E_{XY}(n-1)$ y $E_{XY}(n+1)$, es decir, hay que comparar el valor de la energía asociado a la diada XY con los valores correspondientes a las diadas que se encuentran en la *vecindad inmediata* de la diada XY. Por ejemplo, tomemos la diada 13 de la cadena mostrada en la figura 3.4, y las diadas 12 y 14 que están en su vecindad inmediata:

$$\overset{12}{\overbrace{}}\ \overset{14}{\overbrace{}}$$
$$\ldots 0\ \underbrace{1\ 0}\ 1 \ldots$$

Tenemos que $E_{XY}(12) = E_{01} = 5$, $E_{XY}(13) = E_{10} = 1$, y $E_{XY}(14) = E_{01} = 5$, y por lo tanto $E_{XY}(12) > E_{XY}(13) < E_{XY}(14)$, de donde la diada 13 representa un mínimo en la gráfica de energías. Si, como otro ejemplo, tomamos la diada 8 y las diadas 7 y 9 de su vecindad inmediata, y hacemos el mismo análisis, obtenemos que



$E_{XY}(7) = E_{01} = 5$, $E_{XY}(8) = E_{11} = 9$, $E_{XY}(9) = E_{11} = 9$, de donde $E_{XY}(7) < E_{XY}(8) = E_{XY}(9)$, y por lo tanto, la diada 8 no representa un mínimo en la gráfica de energías.

1 1 1 1	$E_{11} = E_{11} = E_{11}$	no hay mínimo
1 1 1 0	$E_{11} = E_{11} > E_{10}$	no hay mínimo
0 1 1 1	$E_{01} < E_{11} = E_{11}$	no hay mínimo
0 1 1 0	$E_{01} < E_{11} > E_{10}$	no hay mínimo
1 1 0 1	$E_{11} > E_{10} < E_{01}$	**si hay mínimo**
1 1 0 0	$E_{11} > E_{10} < E_{00}$	**si hay mínimo**
0 1 0 1	$E_{01} > E_{10} < E_{01}$	**si hay mínimo**
0 1 0 0	$E_{01} > E_{10} < E_{00}$	**si hay mínimo**
1 0 1 1	$E_{10} < E_{01} < E_{11}$	no hay mínimo
1 0 1 0	$E_{10} < E_{01} > E_{10}$	no hay mínimo
0 0 1 1	$E_{00} > E_{01} < E_{11}$	**si hay mínimo**
0 0 1 0	$E_{00} > E_{01} > E_{10}$	no hay mínimo
1 0 0 1	$E_{10} < E_{00} > E_{01}$	no hay mínimo
1 0 0 0	$E_{10} < E_{00} = E_{00}$	no hay mínimo
0 0 0 1	$E_{00} = E_{00} > E_{01}$	no hay mínimo
0 0 0 0	$E_{00} = E_{00} = E_{00}$	no hay mínimo

Tabla 1

$$E_{00} = E_{11} > E_{01} > E_{10}$$

De los dos ejemplos anteriores vemos que el problema de saber si hay o no un mínimo de energía se reduce a analizar secuencias de cuatro dígitos: en estas secuencias hay que comparar el valor de la energía asociado a la diada "de en medio" con los valores de las energías asociados a las diadas "de las orillas". En el caso binario podemos formar 16 secuencias diferentes de cuatro dígitos cada una, las cuales se muestran en la tabla 1. De esta tabla se observa que sólo 5 de las 16 posibles secuencias de cuatro dígitos conducen a mínimo en la gráfica de energías, y por lo tanto, la probabilidad P_m de tener un mínimo en una cadena binaria aleatoria está dada por

$$P_m = \frac{5}{16}.$$

La anterior es la probabilidad de que en una cadena aleatoria aparesca un mínimo de energía. Si en la cadena hay N diadas, entonces el número N_m de mínimos en dicha cadena será $N_m = P_m N$ aproximadamente. Por lo tanto, la distancia promedio entre los mínimos de energía la podemos calcular como

$$\bar{d} = \frac{N}{N_m} = \frac{N}{P_m N} = \frac{1}{P_m} = \frac{16}{5} = 3.2 \tag{7}$$

lo cual demuestra de forma teórica el resultado (6) obtenido anteriormente.

1 1 1 1	$E_{11} = E_{11} = E_{11}$	no hay mínimo
1 1 1 0	$E_{11} = E_{11} < E_{10}$	no hay mínimo
0 1 1 1	$E_{01} < E_{11} = E_{11}$	no hay mínimo
0 1 1 0	$E_{01} < E_{11} < E_{10}$	no hay mínimo
1 1 0 1	$E_{11} < E_{10} > E_{01}$	no hay mínimo
1 1 0 0	$E_{11} < E_{10} > E_{00}$	no hay mínimo
0 1 0 1	$E_{01} < E_{10} > E_{01}$	no hay mínimo
0 1 0 0	$E_{01} < E_{10} > E_{00}$	no hay mínimo
1 0 1 1	$E_{10} > E_{01} < E_{11}$	**si hay mínimo**
1 0 1 0	$E_{10} > E_{01} < E_{10}$	**si hay mínimo**
0 0 1 1	$E_{00} > E_{01} < E_{11}$	**si hay mínimo**
0 0 1 0	$E_{00} > E_{01} < E_{10}$	**si hay mínimo**
1 0 0 1	$E_{10} > E_{00} > E_{01}$	no hay mínimo
1 0 0 0	$E_{10} > E_{00} = E_{00}$	no hay mínimo
0 0 0 1	$E_{00} = E_{00} > E_{01}$	no hay mínimo
0 0 0 0	$E_{00} = E_{00} = E_{00}$	no hay mínimo

Tabla 2

$$E_{01} < E_{11} = E_{00} < E_{10}$$

Si analizamos los pasos que nos condujeron al resultado (7), vemos que dicho resultado *no* depende de los valores numéricos explícitos (1, 5 y 9) que asignamos a las energías E_{XY} $(X, Y = 1, 0)$. Este resultado depende únicamente de la relación de "... mayor que ...", dada en (3), que existe entre dichas energías . Esperamos, por tanto, que si cambiamos esta relación, cambie también el valor de $\bar{d}$. Y en efecto, la tabla 2 muestra que si la relación de "... mayor que ..." existente entre las E_{XY} fuera

$$E_{01} < E_{00} = E_{11} < E_{10} \tag{8}$$

entonces sólo 4 de las 16 secuencias binarias de cuatro dígitos posibles conducirían a un mínimo en la gráfica de energías, y por lo tanto, la probabilidad P_m de obtener uno de tales mínimos en una cadena binaria aleatoria sería

$$P_m = \frac{4}{16} = \frac{1}{4}$$

lo que se traduce a que, en este caso, la distancia promedio entre mínimos sucesivos a lo largo de la cadena es

$$\bar{d} = 4 \tag{9}$$

Es fácil constatar que este mismo resultado se obtiene si la relación entre las energías asociadas a las diadas de bases fuera

$$E_{00} = E_{11} < E_{10} < E_{01} \tag{10}$$

De cualquier modo, ya sea con la relación (3), o con la (8), o con la (10), podemos decir que para cadenas binarias *aleatorias*, la distancia promedio entre mínimos consecutivos de energía "anda entre 3 y 4". En el apartado siguiente realizaremos este mismo tipo de análisis para secuencias genéticas de ADN compuestas de cuatro símbolos (A, T, C y G).

Mínimos de energía en las secuencias genéticas.

En 1986 K. J. Breslauer y J. Freir *et. al.* publicaron una tabla termodinámica completa de las interacciones entre primeros vecinos de las bases del ADN y del ARN[12]. Reproducimos aquí, en la tabla 3, la tabla de Breslauer–Freir, en la cual se muestran las entalpías (ΔH) y las energías libres (ΔG) de amarre entre parejas de bases complementarias para las 16 diadas que se pueden formar con las cuatro bases del ADN (A, T, C, G), o con las cuatro bases del ARN (A, U, C, G). Dado que dos bases consecutivas y sus complementarias forman un paralelepípedo, las energías de las que se habla se deben entender como un promedio entre las cuatro interacciones resultantes. Por esta razón, sólo 10 de las 16 diadas resultan ser independientes respecto a las energías de amarre asociadas. Por ejemplo,

diada	$\Delta H°(ADN)$	$\Delta G°(ADN)$	$\Delta H°(ARN)$	$\Delta G°(ARN)$
$AA\backslash TT$	9.1	1.9	6.6	0.9
$CC\backslash GG$	11.0	3.1	12.2	2.9
$AC\backslash GT$	6.5	1.3	10.2	2.1
$AG\backslash CT$	7.8	1.6	7.6	1.7
$GA\backslash TC$	5.6	1.6	13.3	2.3
$GT\backslash AC$	6.5	1.3	10.2	2.1
AT	8.6	1.3	5.7	0.9
TA	6.0	1.9	8.1	1.1
CG	11.9	3.6	8.0	2.0
GC	11.1	3.1	14.2	3.4

Tabla 3

Tabla de Breslauer–Freir. Valores absolutos de la entalpía $\Delta H°$ y de la energía libre $\Delta G°$ correspondientes a diadas tanto de ADN como de ARN. Todos los valores se refieren a la ruptura de la interacción en una molécula helicoidal a 1M. NaCl, 25°C y pH7. Las unidades de $\Delta G°$ y $\Delta H°$ son kcal/mol (1 cal = 4.184 J).

las diadas AG y CT tienen la misma energía asociada, ya sea entalpía o energía libre, ya que ambas diadas son complementarias:

$$5' \longrightarrow \underline{A} - \underline{G} \longrightarrow 3'$$
$$3' \longleftarrow \overline{T} - \overline{C} \longleftarrow 5'$$

Por otro lado, debido a la estructura helicoidal de los enlaces complementarios, las energías no son simétricas, es decir, la energía E_{AT}, por ejemplo, no es la misma que la energía E_{TA}. Sucedía algo análogo en el caso binario, en donde teníamos cuatro diadas, pero sólo tres de éstas eran independientes respecto a los valores de energía asociados (he aquí la justificación de las relaciones (3), (8) y (10)). Y al igual que en el caso binario, buscaremos la distancia promedio $\bar{d}$ entre mínimos de energía consecutivos a lo largo de las cadenas genéticas *construidas al azar* y compuestas de cuatro símbolos.

Consideremos por ejemplo las tablas de entalpía del ARN y del ADN. Sabemos que el valor de $\bar{d}$ no depende de los valores numéricos explícitos de las energías asignadas a las diadas, sino únicamente de la relación "... mayor que ..." existente entre dichas energías. Para encontrar la probabilidad P_m de obtener un mínimo de energía a lo largo de la cadena genética, tenemos que analizar secuencias de cuatro bases, comparando el valor de la energía de la diada de "en medio" con los valores de las energías de las diadas de "las

orillas". Por ejemplo, según la tabla de entalpías del ADN (tabla 3) la secuencia

$$\ldots A\ T\ A\ G \ldots$$

conduce a un mínimo en la gráfica de energías, ya que

$$\Delta H^{\circ}_{AT} > \Delta H^{\circ}_{TA} < \Delta H^{\circ}_{AG}.$$

En este caso podemos formar 256 cadenas de cuatro elementos cada una utilizando para ello las cuatro bases del ADN (A, T, C, G), o las cuatro del ARN (A, U, C, G). Analizando de esta manera todas y cada una de las 256 "cuaternas" de bases (no expondremos aquí en detalle el análisis porque, aunque no es complicado, es laborioso estar revisando cada una de las 256 cuaternas del ARN y cada una de las 256 del ADN, pero el procedimiento es el mismo que para el caso binario mostrado en las tablas 1 y 2), llegamos a la conclusión de que para el ARN, 87 de estas cuaternas conducían a mínimos, mientras que para el ADN lo hacían 83. Por lo tanto, las probabilidades P_{mARN} y P_{mADN} de tener un mínimo de energía lo largo de las cadenas de ARN y ADN, respectivamente, son

$$P_{mARN} = \frac{87}{256}$$

$$P_{mADN} = \frac{83}{256}$$

De aquí que las distancias promedio $\bar{d}_{ARN}$ y $\bar{d}_{ADN}$ entre mínimos de energía consecutivos a lo largo de las cadenas de ARN y ADN sean, respectivamente

$$\bar{d}_{ARN} = \frac{256}{87} = 2.94 \tag{11}$$

$$\bar{d}_{ADN} = \frac{256}{83} = 3.08$$

Vemos que los valores numéricos de estas distancias se encuentran peligrosamente próximos al número mágico 3, tanto, que es casi imposible seguir llamándole "mágico" a este número. En la sección siguiente discutiremos las consecuencias físicas de que los mínimos de energía a lo largo de las cadenas genéticas estén separados en promedio cada tres bases. Por el momento quisieramos añadir una observación más acerca de nuestros resultados.

Hemos visto que en el caso binario la distancia promedio $\bar{d}_b$ entre mínimos de energía consecutivos a lo largo de cadenas *aleatorias* era de $\bar{d}_b = 3.2$ o $\bar{d}_b = 4$, dependiendo de si la relación de "... mayor que ..." entre las energías asociadas a las diadas está dada por (3) o por (8), respectivamente. En el caso de cadenas aleatorias construidas con cuatro bases, ya sean de ADN o de ARN, los mínimos de energía están separados, para todos los fines que nos ocupan, por una distancia promedio que podemos tomar como $\bar{d} = 3$, y en este caso, las energías asociadas a las diadas vienen dadas por la tabla de Breslauer, no las podemos cambiar. ¿Este valor de $\bar{d} = 3$ depende de que tanto el ADN como el ARN están construidos con cuatro bases? En otras palabras, ¿hubiera cambiado drásticamente

el valor de $\bar{d}$ si el ADN y el ARN estuvieran compuestos de seis, o de ocho, o de diez bases, y no sólo de cuatro, como de hecho lo están?

diada	$\Delta H°$	diada	$\Delta H°$
$AA\backslash TT$	1	$CX\backslash YG$	12
$CC\backslash GG$	2	$CY\backslash XG$	13
$XX\backslash YY$	3	$GX\backslash YC$	14
$AC\backslash GT$	4	$GY\backslash XC$	15
$AG\backslash CT$	5	TA	16
$AX\backslash YT$	6	AT	17
$AY\backslash XT$	7	CG	18
$TC\backslash GA$	8	GC	19
$TG\backslash CA$	9	XY	20
$TX\backslash YA$	10	YX	21
$TY\backslash XA$	11		

Tabla 4

"Tabla de Breslauer-Freir" correspondiente a un "ADN" compuesto de 6 bases. En este caso tenemos 21 diadas independientes respecto a los valores de "entalpía" asociados. El valor numérico de estas energías de amarre es irrelevante; lo importante es que tales valores obedezcan una relación de "... mayor que ..." entre ellos.

Con el objetivo de obtener una idea de la respuesta a la pregunta anterior, construímos cadenas de "ADN" compuestas de 6 bases: A, T, C, G, X, Y, con enlaces complementarios A-T, C-G, X-Y. Con estas bases hicimos una "tabla de Breslauer" para las 36 diadas de este 6-ADN, mostrada en la tabla 4, y medimos la $\bar{d}_{6ADN}$ numéricamente utilizando 100 cadenas aleatorias, cada una de longitud 100000. El resultado que obtuvimos de estos cálculos fue el siguiente:

$$\bar{d}_{6ADN} = 3.020 \pm 0.037 \tag{12}$$

es decir, incluso para este ADN hipotético construido con 6 bases la distancia promedio entre mínimos de energía consecutivos a lo largo de la cadena también fue, para todos los fines prácticos, 3. Cabe mencionar aquí que el resultado (12) lo obtuvimos no con una sola tabla, sino que para cada una de las 100 cadenas aleatorias que construímos, también construímos una tabla análoga a la tabla 4 de manera aleatoria, y lo que cambiaba de una tabla a otra era únicamente la relación de "... mayor que ..." existente entre las energías asociadas a las diadas. De modo que el resultado (12) no sólo representa el promedio de 100 cadenas diferentes con la misma tabla, sino de 100 cadenas diferentes con 100 tablas diferentes también.

Y fuimos más lejos aún, haciendo los mismos cálculos para cadenas formadas con 8 bases, A, T, C, G, U, V, X, Y, con enlaces complementarios A-T, C-G, U-V, X-Y. En este caso de un "ADN" construido con 8 bases no mostramos ya la "tabla de Breslauer" correspondiente, pero es análoga a la tabla 4. Utilizando también 100 cadenas aleatorias, cada una de longitud 100000, así como 100 "tablas de Breslauer" diferentes, obtuvimos el resultado ya poco inesperado

$$\bar{d}_{8ADN} = 3.000 \pm 0.005 \tag{13}$$

Vemos de los resultados (7), (9), (11), (12) y (13) que entre mayor es el número de bases con las que construímos las cadenas de "ADN", la distancia promedio entre mínimos de energía consecutivos a lo largo de las cadenas, se va aproximando cada vez más a 3, y de hecho, exceptuando el caso binario en donde se obtiene una $\bar{d}_b = 4$, podemos considerar que el periodo 3 es independiente del número de bases con las cuales se construye el ADN.

El periodo 3 en el origen de la vida

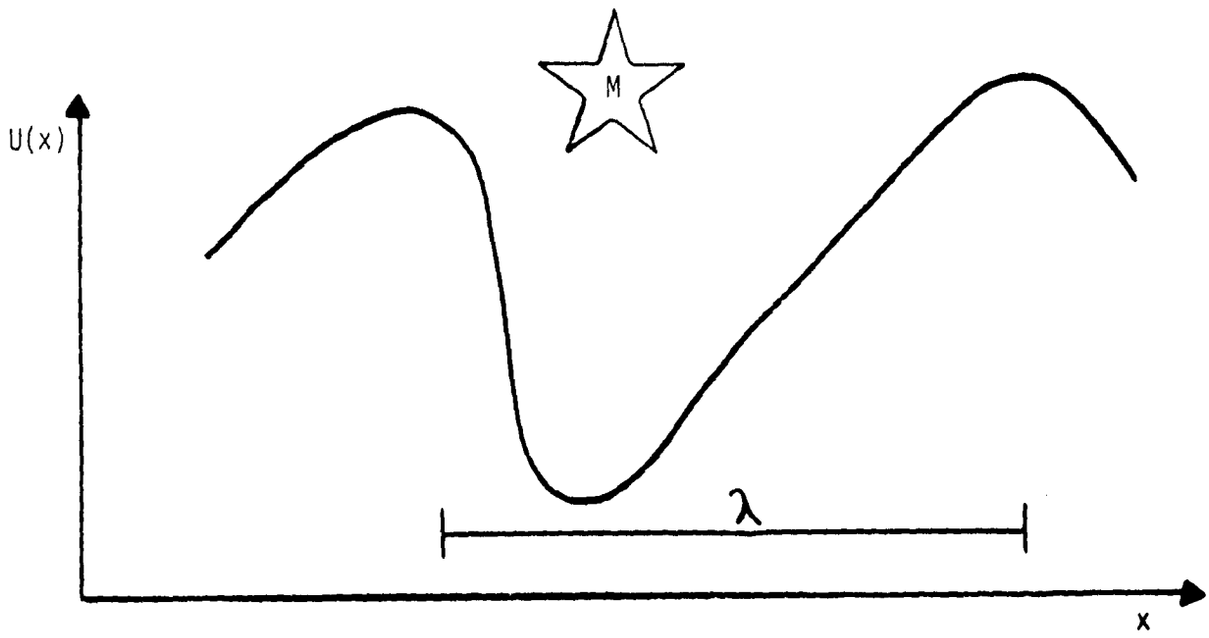
Rcientemente Marcelo O. Magnasco y Mark M. Millonas *et al.* [13],[14] han hecho trabajos que apuntan en la dirección siguiente: han demostrado que si una molécula M está sujeta a un potencial unidimensional $U(x)$ el cual es periódico y asimétrico respecto a sus extremos en un periodo, como se muestra en la figura 3.5, y si además el sistema está sujeto a fluctuaciones aleatorias que no sean ruido blanco gaussiano, entonces la molécula M se moverá en una dirección preferente (ya sea a la "izquierda" o a la "derecha") a lo largo del potencial. Cabe mencionar que tales sistemas son rectificadores de ruido, y son de gran interés en el entendimiento de la maquinaria que opera en "el régimen Browniano", es decir, en escalas muy pequeñas en donde las fluctuaciones juegan el papel más importante. Los ejemplos más directos son los sistemás biológicos, y en particular, el que nos ocupa a nosotros: las secuencias genéticas y el periodo 3.

Si $f(t)$ representa una función de ruido en el tiempo, el que $f(t)$ sea un ruido branco gaussiano significa que

$$\langle f(t)f(t')\rangle = \delta(t-t') \tag{14}$$

donde "$\langle\cdots\rangle$" significa promedio temporal. En los trabajos de Millonas y Magnasco es indispensable que $f(t)$ no sea un ruido balnco gaussiano, es decir, que la condición (14) no se cumpla, porque la probabilidad J de que la molécula viaje a través del potencial $U(x)$ (o lo que es lo mismo, la densidad de corriente, si es que tenemos un flujo de moléculas), es proporcional a $exp[-\langle[f(t)]^2\rangle]$:

$$J \propto \exp^{-\langle[f(t)]^2\rangle} \tag{15}$$



Fugura 3.5

Potencial típico que permite el desplazamiento de la molécula M en una dirección preferente a lo largo del potencial cuando el sistema está sometido a un ruido "coloreado". λ es el periodo del potencial.

y vemos de las ecuaciones (14) y (15) que si $t = t'$ entonces $J = 0$ y no se tedría ningún flujo de moléculas. Por eso se pide que $f(t)$ sea un ruido "coloreado" y no simplemente un ruido térmico.

Hemos visto en este capítulo que los mínimos de energía asociados a diadas de bases se encuentran, a lo largo de las secuencias genéticas, separados en promedio cada tres bases. Esto produciría, a grosso modo, un potencial cuasiperiódico del tipo indicado en la figura 3.5, donde en este caso $\lambda \approx 3$. Compaginando esto último con los trabajos de Millonas y Magnasco, es posible dar tentativamente una explicación a la naturaleza del código genético: la información en el mARN se lee por codones (triadas de bases) porque los mínimos de energía a lo largo de esta molécula están espaciados cada tres bases. La molécula M de la figura 4 puede ser un mARN, y la molécula que genera el potencial un rARN. Por tanto, es lógico (aunque no necesariamente cierto) pensar que la "maquinaria celular" evolucionó de tal forma que el mARN se vaya moviendo de tres en tres a lo largo del rARN, precisamente porque cada tres bases hay un mínimo de energía. Es importante dejar claro que no estamos tratando de explicar el complejo mecanismo de síntesis de proteínas tal y como lo conocemos actualmente. Mas bién, lo que queremos hacer es dar una explicación de cómo se *originó* este mecanismo al inicio de la vida. Seguramente las primeras moléculas de ARN que aparecieron fueron pequeñas y tenían sus nucleótidos

ordenados al azar (por eso nosotros trabajamos con secuencias al azar). El que los mínimos de energía consecutivos aparecieran separados cada tres bases a lo largo de estas moléculas pudo ser el punto de partida que aprovechó la Evolución para dar lugar a los complejos mecanismos biológicos involucrados en la síntesis de proteínas que ahora conocemos.

Además, según hemos visto, el periodo 3 es independiente del número de bases con las cuales se construyan las secuencias genéticas, exceptuando el caso binario en donde se tiene un periodo 4. Por tanto, basta con que la Naturaleza hubiese dispuesto de más de dos bases para construir a las moléculas portadoras de la información genética y que estas bases hayan seguido un princípio de complementariedad (que es lu *único* que hemos supuesto en este trabajo), para que el periodo 3, lejos de ser la excepción, sea la regla. Y volviendo a la respuesta que mencionamos al princípio que se le había dado a la pregunta con la cual comenzamos este capítulo, en base a lo que hemos visto creemos que no fue la información contenida en los ácidos nucleicos la que se tuvo que adaptar a los 20 aminoácidos ya existentes, sino que la evolución explotó al máximo esta característica genérica del periodo 3 en las secuencias genéticas construídas con más de dos bases, para "construir" a la maquinaria celular sintetizadora de proteínas, lo cual por lo demás, está de acuerdo con las ideas actuales que tienen los biólogos evolutivos alrededor de que las primeras moléculas biológicas que aparecieron fueron de ARN.

Finalmente, cabe puntualizar que para secuencias genéticas reales, no solamente es el promedio de la distancia entre mínimos de energía el que cae en tres, sino que también cae en tres el valor más probable de estas distancias, tal y como lo demuestran los histogramas que aparecen en el apéndice de este trabajo, lo cual nos da más bases aún para sospechar que la Naturaleza explotó al máximo esta característica de periodo tres que presentan las secuencias genéticas.

Capítulo 4

Matrices de correlación y mapas de correlación genética

La Física y la Química describen las leyes de interacción de la materia, pero ésta no sólo interactúa, sino que también se organiza.

Albert Lehninger.

Variabilidad del VIH.

Resulta ser innegable el hecho de que el índice de mutación del VIH es muy elevado: la retrotranscriptasa que usa el virus para retrotranscribir al ARN viral a una nueva cadena de ADN (provirus), comete, en promedio, un error cada 2000 nucleótidos incorporados, aproximadamente. De tamaña impresición, o infidelidad, se desprende la notable capacidad del virus para oponer resistencia a drogas de diversa índole: no paran de generarse nuevas formas de proteínas víricas en el curso de una infección. Más aún, se sabe ahora que existen regiones del genoma viral, llamadas *hot spots*, y que son aquellas partes del gen *env* que codifican para la proteína externa gp120, en donde el nivel de error que comete la retrotranscriptasa alcanza a 1 de cada 70 nucleótidos incorporados. Lo anterior ha dado como resultado que se hayan encontrado virus, provenientes de un mismo paciente, cuyas secuencias completas de nucleótidos difieren hasta en un 10%.

Si comparamos lo anterior con los índices de mutación presentes en la replicación del ADN de las células eucariotes, en donde los errores se cometen con una frecuencia de 1 en 10^9–10^{12}, vemos que la variabilidad del virus del SIDA es al menos un millón de veces más grande que la de los organismos celulares eucariotes. Tal variabilidad del VIH llevó a Manfred Eigen[15] a clasificar al virus del SIDA, ya no como una especie, sino como una *cuasiespecie*, es decir, como una familia de organismos cuyos genomas son bastante diferentes unos de otros, pero que sin embargo, conservan los rasgos estructurales que le confieren su identidad al virus del SIDA como tal.

La idea de Eigen de la cuasiespecie tiene que ver, de hecho, con la siguiente pregunta:

¿hasta dónde llega la variabilidad del virus? En otras palabras, si bién es cierto que el virus muta demasiado, también lo es el que , a pesar de las mutaciones, el virus sigue siendo *el* virus del SIDA. Pese a las mutaciones, el virus no se transforma en otro virus que cause otra enfermedad que no sea SIDA. Esto sugiere fuertemente que en el genoma viral *deben* de existir regiones conservadas, regiones que no cambien y que son las responsables de que el virus del SIDA siga siendo, como ya lo hemos puntualizado, *el* virus del SIDA. En este capítulo presentaremos métodos de análisis que revelan la presencia de una estructura general, conservada, en los genomas virales del VIH, estructura que le confiere su identidad al virus del SIDA como tal.

Matriz de correlación.

Consideremos a las secuencias genéticas como simples cadenas formadas por cuatro tipos diferentes de letras (A, T, C, G), tal y como lo muestra la figura 3.1. Definimos el índice de correlación $C_{\alpha\beta}(d)$ entre las bases α y β como

$$C_{\alpha\beta}(d) = \frac{N_{\alpha\beta}(d)N - N_\alpha N_\beta}{N^2} \qquad (4.1)$$
$$\alpha, \beta = A, T, C, G.$$

donde $N_{\alpha\beta}(d)$ es el número de parejas $\alpha\beta$ que hay a lo largo de toda la cadena, N_α es el número de bases α, N_β el número de bases β, y N es el número total de bases en la cadena; d es la distancia que hay entre la base α y la base β. Por ejemplo, consideremos la siguiente secuencia:

primeros vecinos

$$A\,T\,C\,A\,T\,C\,T\,T\,G\,G\,A\,A\,T\,T\,C\,T\,G\,C\,C\,A\,C\,A\,T\,T\,C\,C\,A\,T.$$

terceros vecinos

La cadena tiene 28 bases en total, y por lo tanto

$$N = 28.$$

Hay 7 A's y 10 T's, de modo que

$$N_A = 7$$

$$N_T = 10.$$

Si tomamos $d = 1$ (primeros vecinos), en la cadena existen 5 parejas AT a primeros vecinos, de aquí que

$$N_{AT}(1) = 5$$

y por lo tanto, $C_{AT}(1)$ queda como

$$C_{AT}(1) = \frac{N_{AT}(1)N - N_A N_T}{N^2}$$
$$C_{AT}(1) = \frac{(5)(28) - (7)(10)}{28^2}$$
$$C_{AT}(1) = \frac{5}{56}$$

Por otro lado, si $d = 3$ (terceros vecinos), vemos que en la cadena sólo hay 3 parejas AT a terceros vecinos, de modo que

$$N_{AT} = 3$$

y consecuentemente

$$C_{AT}(3) = \frac{(3)(28) - (7)(10)}{28^2} = \frac{1}{56}$$

Con estos dos ejemplos basta para saber como opera el índice de correlación $C_{\alpha\beta}(d)$. ¿Cuál es el significado de este índice de correlación? Para contestar la pregunta reescribamos la ecuación (4.1) como

$$C_{\alpha\beta}(d) = \frac{N_{\alpha\beta}(d)}{N} - \left(\frac{N_\alpha}{N}\right)\left(\frac{N_\beta}{N}\right) \tag{4.2}$$

El primer término, $N_{\alpha\beta}(d)/N$, no es más que el estimador de la probabilidad de que aparezca la pareja $\alpha\beta$ a lo largo de la cadena, estando las bases α y β separadas una distancia d. El segundo término que aparece en el miembro derecho de la ecuación (4.2) está compuesto por dos factores: N_α/N es el estimador de la probalilidad de que la base α aparezca en la cadena, mientras que N_β/N es el correspondiente estimador referente a la base β. Por lo tanto, el producto

$$\left(\frac{N_\alpha}{N}\right)\left(\frac{N_\beta}{N}\right)$$

no es más que la probabilidad empírica de tener a la pareja $\alpha\beta$ en la cadena, si ésta estuviera hecha al azar. Vemos pues que el índice de correlación $C_{\alpha\beta}(d)$ es el estimador de la probabilidad de que en la cadena tengamos a la pareja $\alpha\beta$, *restando el azar*, es decir, quitando la probabilidad de que dicha pareja apareciera si la cadena estuviese construída aleatoriamente.

Luego, el índice de correlación $C_{\alpha\beta}(d)$ nos da la probabilidad (restando el azar) de que a una distancia d de la base α, se encuentre la base β. Sin embargo, es necesario decir que esta correlación es *estadística*, y no debe entenderse como una correlación de "causa-efecto". Una correlación estadística de largo alcance en las secuencias genéticas

de ADN significa que la ocurrencia o densidad de bases tiende a variar de alguna forma regular (si es que hay correlación) en regiones separadas por grandes distancias. Esto es un concepto completamente diferente al que se tiene cuando uno dice "que una región de la secuencia genética del ADN tiene una influencia biológica directa sobre otra región que está alejada".

En los ejemplos que mostramos anteriormente para ilustrar como "funciona" el índice de correlación, tomamos a la pareja AT. Pero igual hubieramos podido tomar a la pareja AC, o a la GG. En realidad, podemos formar hasta 16 parejas utilizando las cuatro bases del ADN. Esto da como consecuencia que tengamos 16 tipos distintos de índices de correlación $C_{\alpha\beta}$: cuatro valores para el índice α y otros cuatro para el índice β. Lo que tenemos es entonces, una matriz de 4×4 cuyos elementos son los índices de correlación $C_{\alpha\beta}$, y si hacemos la asociación de índices

$$\begin{aligned} A &\longrightarrow 1 \\ T &\longrightarrow 2 \\ C &\longrightarrow 3 \\ G &\longrightarrow 4 \end{aligned} \tag{4.3}$$

podemos escribir esta matriz de correlación como

$$\mathbf{C}(d) = \begin{pmatrix} C_{11}(d) & C_{12}(d) & C_{13}(d) & C_{14}(d) \\ C_{21}(d) & C_{22}(d) & C_{23}(d) & C_{24}(d) \\ C_{31}(d) & C_{32}(d) & C_{33}(d) & C_{34}(d) \\ C_{41}(d) & C_{42}(d) & C_{43}(d) & C_{44}(d) \end{pmatrix} \tag{4.4}$$

donde hemos puesto explícitamente la dependencia en d para enfatizar que esta matriz depende de la distancia de correlación: a primeros vecinos tenemos una matriz $\mathbf{C}(1)$, a segundos vecinos tenemos otra matriz $\mathbf{C}(2)$, a terceros vecinos ... etc.

Realmente, aunque la matriz $\mathbf{C}(d)$ aparece en (4.4) como una matriz de 4×4, sólo 9 de sus 16 entradas son independientes. Lo anterior es consecuencia de que los elementos de dicha matriz cumplen con las restricciones:

$$\sum_{\alpha=1}^{4} C_{\alpha\beta}(d) = 0 \qquad \beta = 1, 2, 3, 4 \tag{4.5a}$$

$$\sum_{\beta=1}^{4} C_{\alpha\beta}(d) = 0 \qquad \alpha = 1, 2, 3, 4 \tag{4.5b}$$

es decir, la suma de los elementos de un renglón cualquiera de la matriz es nula, y también lo es la suma de los elementos de cualquier columna. Que las restricciones (4.5) se cumplen, podemos demostrarlo a partir de la definición del índice de correlación (4.1) o (4.2). Demostremos, por ejemplo, la relación (4.5a) utilizando el valor $\beta = 1$ (en los siguientes cálculos obviamos la dependencia en d):

$$\sum_{\alpha=1}^{4} C_{\alpha 1} = C_{11} + C_{21} + C_{31} + C_{41}$$

o bien, utilizando la asociación de índices dada en (4.3), tenemos

$$\begin{aligned}\sum_{\alpha=1}^{4} C_{\alpha 1} &= C_{AA} + C_{TA} + C_{CA} + C_{GA} \\ &= \left[\frac{N_{AA}}{N} - \frac{N_A}{N}\frac{N_A}{N}\right] + \left[\frac{N_{TA}}{N} - \frac{N_T}{N}\frac{N_A}{N}\right] \\ &+ \left[\frac{N_{CA}}{N} - \frac{N_C}{N}\frac{N_A}{N}\right] + \left[\frac{N_{GA}}{N} - \frac{N_G}{N}\frac{N_A}{N}\right]\end{aligned}$$

y reagrupando términos tenemos

$$\sum_{\alpha=1}^{4} C_{\alpha 1} = \frac{1}{N}\left[N_{AA} + N_{TA} + N_{CA} + N_{GA}\right] - \frac{N_A}{N^2}\left[N_A + N_T + N_C + N_G\right] \tag{4.6}$$

Ahora bien, lo que aparece encerrado entre corchetes en el primer término de la ecuación anterior, es el número total de parejas AA que hay en la cadena, más el número de parejas TA, más el número de parejas CA, más el número de parejas GA. ¡Pero esto no es más que N_A, el número total de A's que hay en la cadena! Lo anterior es porque estamos agotando las posibilidades de parejas que tengan una A: antes de una A siempre hay, o bien otra A (AA), o una T (TA), o una C (CA), o una G (GA) y nada más. Por tanto, la suma que aparece entre corchetes en el primer término de la ecuación (4.6) nos da el número total de parejas en la cadena que tengan una A, lo cual es igual al número total de A's en la cadena:

$$N_{AA} + N_{TA} + N_{CA} + N_{GA} = N_A \tag{4.7}$$

Por otro lado, es claro que

$$N_A + N_T + N_C + N_G = N \tag{4.8}$$

es decir, el número total de A's más el número total de T's más el número total de C's más el de G's nos da el número total de bases en la cadena. Sustituyendo los resultados (4.7) y (4.8) en la ecuación (4.6) obtenemos

$$\begin{aligned}\sum_{\alpha=1}^{4} C_{\alpha 1} &= \frac{1}{N}\left[N_A\right] - \frac{N_A}{N^2}\left[N\right] \\ &= \frac{N_A}{N} - \frac{N_A}{N} = 0\end{aligned}$$

que es lo que queríamos demostrar. Análogamente se demuestran el resto de las ecuaciones de restricción (4.5). Nótese además que sólo siete de estas ocho ecuaciones de restricción son independientes.

Mapas de correlación genética.

Cuando se trabaja con matrices, normalmente uno lo que quiere es trabajar con las partes *invariantes* de la matriz. Por ejemplo, Kunihiko Kaneko *et al.* [16]han trabajado con la función de información mutua, $M(d)$, definida como

$$M(d) = \sum_{\alpha=1}^{4}\sum_{\beta=1}^{4} \frac{N_{\alpha\beta}(d)}{N} \log\Big(\frac{NN_{\alpha\beta}(d)}{N_\alpha N_\beta}\Big)$$

y han definido la función covarianza, o función de autocorrelación, como

$$\begin{aligned} cov(d) &= \sum_{\alpha=1}^{4}\Big(\frac{N_{\alpha\alpha}(d)}{N} - \frac{N_\alpha^2}{N^2}\Big) \\ &= Tr[\mathbf{C}] \end{aligned}$$

donde $Tr[\mathbf{C}]$ indíca la traza de la Matriz $\mathbf{C}$, la cual es un invariante de dicha matriz. La función $M(d)$ fue introducida primero en el marco de la teoría de la información, y recientemente se ha aplicado al estudio de sistemas dinámicos caóticos. La función de información mutua se considera ahora como una medida estandard de la correlación, y aunque nosotros no la vamos a utilizar en nuestro trabajo, la mencionamos para dejar sentado que técnicas de análisis semejantes (pero no iguales) a las que hemos desarrollado, han sido aplicadas por otros autores y en otros contextos.

Lo que nosotros hizimos fue *diagonalizar* a la matriz $\mathbf{C}(d)$ y trabajar con sus *eigenvalores*. Recordamos aquí que los eigenvalores λ de la matriz $\mathbf{C}$ los obtenemos resolviendo el polinomio característico

$$det|\mathbf{C} - \lambda\mathbf{I}| = 0 \tag{4.9}$$

siendo $\mathbf{I}$ la matriz identidad. En el caso general de una matriz de 4×4, el polinomio característico es un polinomio de cuarto grado, por lo tanto, tendrá cuatro raíces: λ_1, λ_2, λ_3, y λ_4. Es decir, en general, una matriz de 4×4 tiene cuatro eigenvalores. Sin embargo, debido a las restricciones (4.5) que presenta la matriz de correlación $\mathbf{C}$, tenemos que sólo tres de las cuatro columnas de la matriz son independientes, y sólo tres de los cuatro renglones de la matriz son también independientes. Lo anterior trae como consecuencia que uno de los eigenvalores de nuestra matriz de correlación sea siempre nulo. Esto podemos demostrarlo fácilmente si en el polinimio característico dado en (4.9) introducimos las restricciones (4.5). Haciendo lo anterior, obtenemos que el polinomio característico de $\mathbf{C}$ se reduce a:

$$\begin{aligned} \lambda\Big[&\lambda^3 - (C_{11}+C_{22}+C_{33}+C_{44})\lambda^2 \\ &+ (C_{11}C_{22}+C_{11}C_{33}+C_{11}C_{44}+C_{22}C_{33}+C_{22}C_{44}+C_{33}C_{44} \\ &- C_{12}C_{21}-C_{13}C_{31}-C_{14}C_{41}-C_{23}C_{32}-C_{24}C_{42}-C_{34}C_{43})\lambda \\ &+ 4(C_{11}C_{23}C_{32}+C_{22}C_{13}C_{31}+C_{33}C_{12}C_{21} \\ &- C_{12}C_{23}C_{31}-C_{13}C_{21}C_{32}-C_{11}C_{22}C_{33})\Big] = 0 \end{aligned} \tag{4.10}$$

Vemos de (4.10) que efectivamente, uno de los eigenvalores, digamos λ_4, es siempre igual a cero. Los otros tres eigenvalores, λ_1, λ_2, y λ_3, son las raíces del polinomio de grado 3 que aparece entre corchetes en (4.10). Debemos hacer notar que como la matriz de correlación $\mathbf{C}(d)$ es función de la distancia d, entonces los eigenvalores también van a ser funciones de d:

$$\lambda_1 = \lambda_1(d)$$
$$\lambda_2 = \lambda_2(d)$$
$$\lambda_3 = \lambda_3(d)$$

Ahora bien, en el caso general, hay dos posibilidades para las raíces de un polinomio de grado tres: las tres raíces son reales, o bien, una es real y las otras dos son complejas conjugadas una de otra. En el caso que nos ocupa de la matriz de correlación $\mathbf{C}(d)$ se presentaron las dos posibilidades mencionadas anteriormente, es decir, para ciertos valores de d los tres eigenvalores $\lambda_1(d)$, $\lambda_2(d)$, y $\lambda_3(d)$ eran reales, y para otros valores de d, teníamos un eigenvalor real y los otros dos eran complejos conjugados uno de otro. Con el objeto de tener una representación gráfica de los eigenvalores como función de la distancia d procedimos de la manera siguiente:

Los tres eigenvalores reales

En el caso de que $\lambda_1(d)$, $\lambda_2(d)$, y $\lambda_3(d)$ sean todos reales, formamos al vector $\vec{v}(d)$ como un vector con dos componentes, la primera de las cuales es la suma de los tres eigenvalores, y la segunda es cero:

$$\begin{aligned} \vec{v}(d) &= (\lambda_1(d) + \lambda_2(d) + \lambda_3(d)\ ,\ 0) \\ &= (Tr[\mathbf{C}(d)]\ ,\ 0) \end{aligned} \tag{4.11}$$

donde hemos utilizado el hecho de que $Tr[\mathbf{C}(d)] = \lambda_1(d) + \lambda_2(d) + \lambda_3(d)$ —$Tr[\mathbf{C}]$ el la *traza* de la matriz $\mathbf{C}$.

Un eigenvalor real y dos complejos conjugados

Supongamos que $\lambda_1(d)$ es el eigenvalor real, y $\lambda_2(d)$ y $\lambda_3(d)$ son los eigenvalores complejos conjugados. Podemos entonces escribir a $\lambda_2(d)$ y $\lambda_3(d)$ como

$$\lambda_2 = a + ib$$
$$\lambda_3 = a - ib$$

siendo b la parte imaginaria de los eigenvelores. Por conveniencia, tomamos a b como un número positivo.

En este caso, al vector $\vec{v}(d)$ lo construímos como

$$\begin{aligned} \vec{v}(d) &= (\lambda_1(d) + \lambda_2(d) + \lambda_3(d)\ ,\ b) \\ &= (Tr[\mathbf{C}(d)]\ ,\ b) \end{aligned} \tag{4.12}$$

es decir, ahora estamos tomando a la "componente y" del vector $\vec{v}(d)$ como la parte imaginaria positiva del eigenvalor complejo.

El objetivo de definir un vector $\vec{v}(d)$ a través de las definiciones (4.11) y (4.12) es el de que cuando obtengamos tres eigenvalores reales, los representemos por medio de un vector "horizontal" en el plano, mientras que cuando obtengamos eigenvalores complejos, los representemos con un vector que tiene cierta inclinación respecto a la horizontal.

Ya que construímos al "vector de eigenvalores" $\vec{v}(d)$ (así le llamamos a este vector) de acuerdo con las definiciones (4.11) y (4.12), según sea el caso, calculamos $\vec{v}(1)$, $\vec{v}(2)$, $\vec{v}(3), \ldots,$ $\vec{v}(d_{max})$ para distancias de correlación que fueran desde $d = 1$ hasta una cierta $d = d_{max}$, siendo ésta la distancia máxima a la cual estamos calculando correlaciones de parejas de bases a lo largo de la secuencia genética. De un modo un tanto arbitrario, tomamos $d_{max} \leq 0.2N$, siendo N el número total de bases en dicha secuencia. La razón de la cota superior de $0.2N$ impuesta a d_{max} estriba en lo siguiente: para que la "longitud efectiva" de la secuencia genética sobre la cual estamos midiendo las correlaciones sea siempre igual a N, es necesario "cerrar" la secuencia, esto es, unir el final de tal secuencia con el comienzo. Por ejemplo, si $N = 100$ y $d = 1$, entonces para calcular los índices de correlación $C_{\alpha\beta}$ podemos contar a las parejas que, a lo largo de la secuencia, se encuentran desde la posición 1 hasta la posición 99 inclusive. Sin embargo, si $N = 100$ y $d = 50$, sólo podíamos contar a las parejas cuya primera base se encuentra desde la posición 1 hasta la posición 50. En el primer caso, la "longitud efectiva" de la secuencia sobre la cual estamos calculando las correlaciones sería de 99, y en el segundo caso sería de 50, aunque la longitud de la secuencia genética *total* sea de 100 bases. Por este motivo tenemos que "cerrar" la secuencia correlacionando "la punta final" con "la punta inicial". No obstante, no hay motivo para que estas "puntas" estén correlacionadas, por lo cual, acordamos no tomar casos extremos como el que vimos en el ejemplo anterior, donde $d = N/2$. De aquí que hayamos tomado como cota superior para d a $0.2N$, con lo cual garantizamos que a lo más, estamos calculando correlaciones entre la décima parte final de la cadena con la décima parte inicial.

Volviendo al vector de eigenvalores $\vec{v}(d)$, una vez que tenemos $\vec{v}(1)$, $\vec{v}(2)$, $\vec{v}(3)$, ..., $\vec{v}(d_{max})$, los sumamos vectorialmente construyendo una curva formada por todos estos vectores, tal y como se muestra en la figura 4.1. A este tipo de curvas les llamaremos *mapas de correlación genética* (o a veces simplemente "mapas") para referirnos a ellas de alguna manera.

Debemos enfatizar que cada vector $\vec{v}(d)$, o bien, cada punto del mapa de correlación genética, representa un aspecto *global* de la secuencia genética. El vector de eigenvalores $\vec{v}(1)$ está asociado a la matriz de correlaciones entre parejas de bases a primeros vecinos a lo largo de *toda* la secuencia; el vector de eigenvalores $\vec{v}(2)$ está asociado a la matriz de correlaciones de parejas de bases a segundos vecinos a lo largo de *toda* la secuencia genética, etc. Por esta razón, no debemos asociar el princípio del mapa de correlación genética con el princípio de la secuencia genética, o la parte final de dicho mapa con la parte final de la secuencia, o cosas por el estilo. Valga la pena insistir en que cada punto

del mapa representa una correlación entre bases a lo largo de **toda la secuencia.**

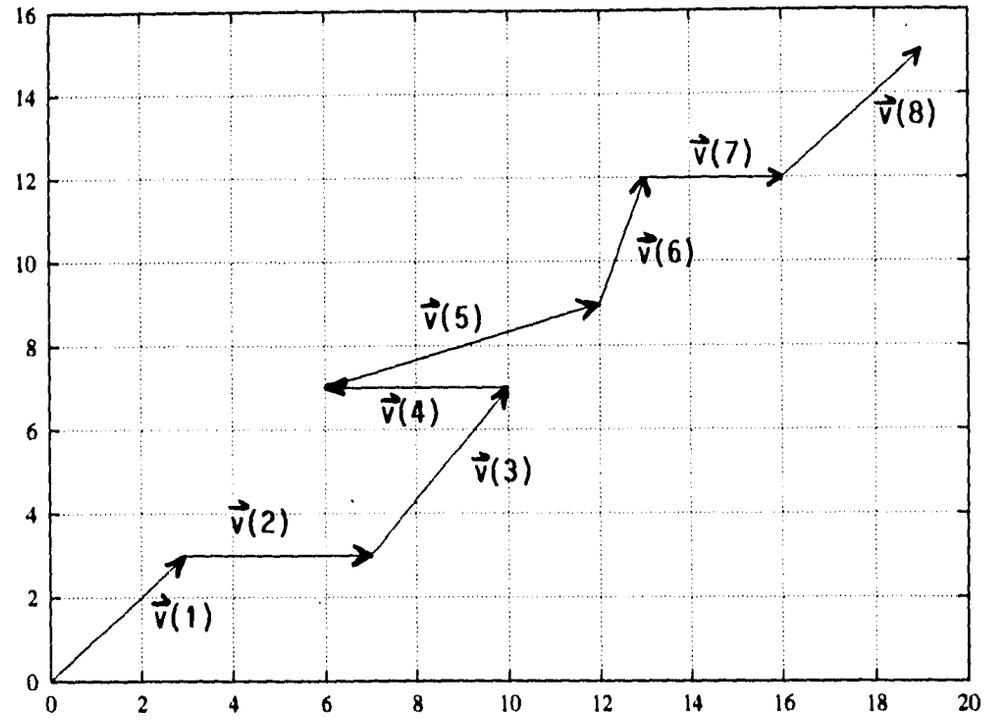


Figura 4.1

Manera en que construímos los mapas de correlación genética a partir de los vectores de eigenvalores.

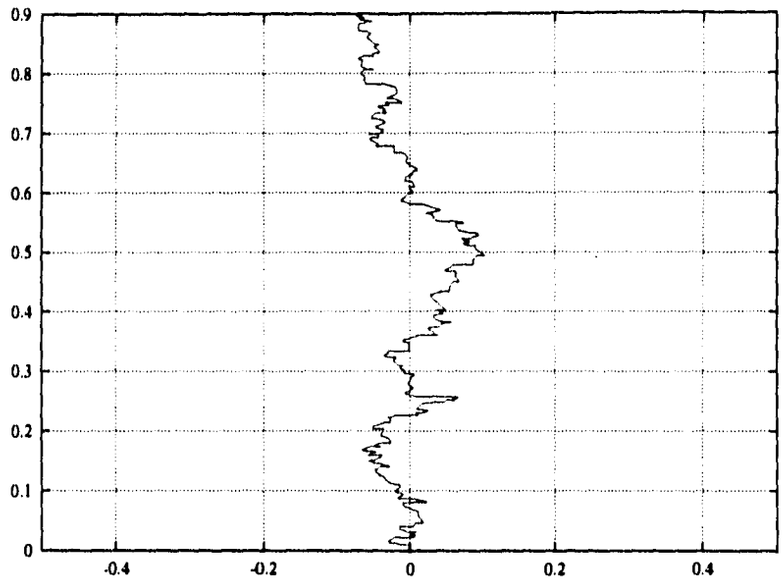
Hasta aquí, lo que hemos hecho es presentar las herramientas que utilizaremos para analizar las secuencias genéticas; lo que resta por hacer ahora, es aplicarlas. Comenzaremos con los casos usuales: una secuencia aleatoria generada con la computadora, y una secuencia completamente periódica, cada una de 2500 bases de longitud. La secuencia periódica seguía el patrón 100 A's, 100 T's, 100 C's, 100 G's, 100 A's, 100 T's, 100 C's, 100 G's, 100 A's, ..., etc., así hasta 2500:

$$\underbrace{AAA\ldots A}_{100veces}\underbrace{TTT\ldots T}_{100veces}\underbrace{CCC\ldots C}_{100veces}\underbrace{GGG\ldots G}_{100veces}\underbrace{AAA\ldots A}_{100veces}\ldots$$

En la figura 4.2(a) y 4.2(b) se muestran los mapas de correlación genética correspondientes a las secuencias aleatoria y periódica, respectivamente, para distancias de correlación desde $d = 1$ hasta $d_{max} = 500$ en ambos mapas. Notemos la diferencia de escalas que existe entre el mapa correspondiente a la secuencia aleatoria y el mapa correspondiente a la secuencia periódica: las escalas en el primer mapa son mucho menores que las escalas en el segundo mapa, lo cual es completamente natural si recordamos que definimos al índice de correlación $C_{\alpha\beta}$ como la probabilidad $P_{\alpha\beta}$ de que aparezca la pareja $\alpha\beta$ en la cadena, menos la probabilidad $P_\alpha P_\beta$ de que esta pareja aparezca si la cadena estuviera hecha

aleatoriamente. Pero si efectivamente la cadena es aleatoria, entonces $P_{\alpha\beta} = P_\alpha P_\beta$, y por lo tanto, $C_{\alpha\beta} = 0$ (esto es en el caso teórico; numéricamente se tiene $C_{\alpha\beta} \simeq 0$). Esto explica por qué el mapa de la figura 4.2(a) tiene escalas tan pequeñas comparadas con el de la figura 4.2(b), que corresponde a una secuencia periódica altamente correlacionada.

(a)



(b)

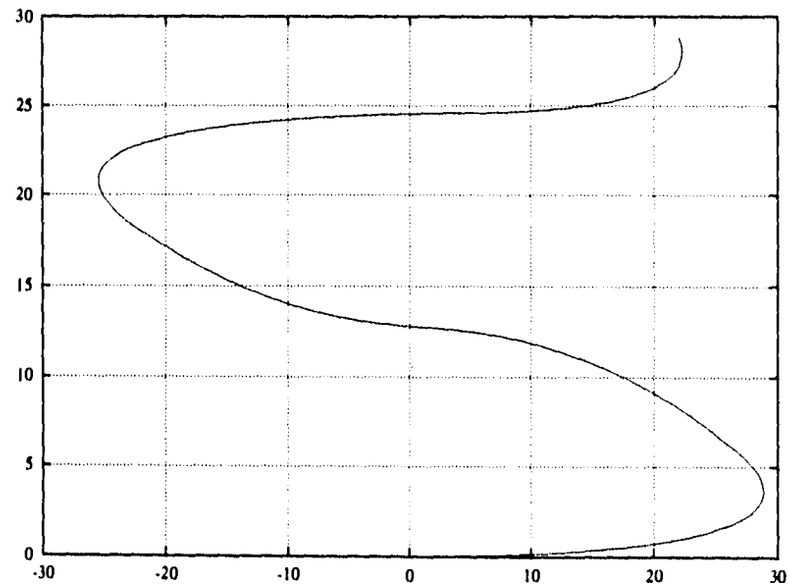
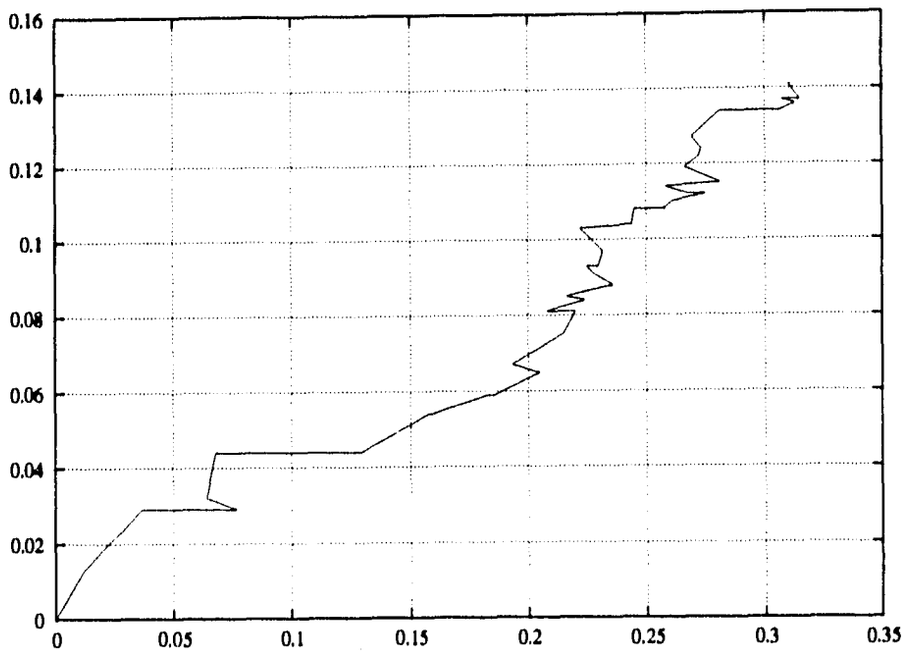


Figura 4.2

(a) Mapa de correlación genética correspondiente a una secuencia aleatoria de 2500 bases de longitud. (b) Mapa correspondiente a una secuencia completamente periódica (ver texto) también de 2500 bases de longitud. Ambos mapas están calculados con $d_{max} = 500$.

(a)



(b)

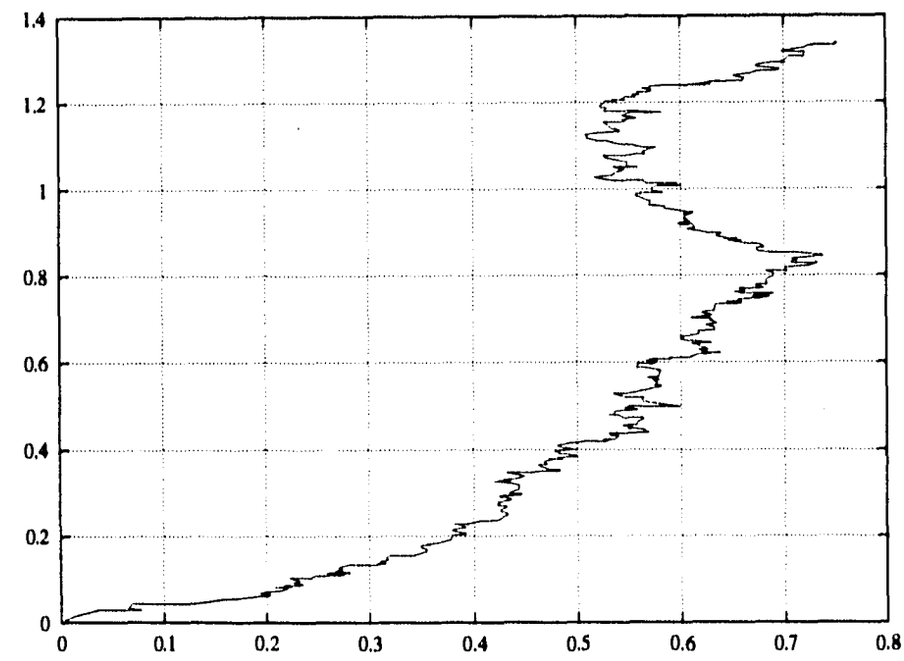


Figura 4.3

Mapas de correlación genética correspondientes al gen env de un virus de SIDA tipo 1. (a) $d_{max} = 50$ *y* (b) $d_{max} = 500$.

En la figura 4.3(a) mostramos el mapa genético correspondiente a la secuencia genética del gen *env* de un virus de sida tipo 1 (vease la figura 3.1) para $d_{max} = 50$ y en la figura 4.3(b) mostramos el mapa correspondiente a la misma secuencia, pero con $d_{max} = 500$. Comparando estos mapas con los dados en la figura 4.2, vemos que la estructura de la secuencia genétca del gen *env* de este virus en particular, está lejos de ser azarosa, pero tampoco está altamente correlacionada.

Realmente, esto es lo más que podemos decir acerca de la estructura del gen *env* a partir de la figura 4.3 solamente. Sin embargo, recordemos que al princípio de este capítulo mencionábamos que el gen *env* es la parte más variable de todo el genoma del VIH. Por lo tanto, lo que nos interesa es comparar los mapas de correlación genética correspondientes al gen *env* de *diferentes* VIH. En otras palabras, si el gen *env*, que es la parte más variable del genoma del VIH, cambiase tanto de un virus a otro que su variabilidad pudiera considerarse como "casi" aleatoria, por decirlo de alguna forma, es de esperarse que el mapa de correlación genética de este gen también cambiara sustancialmente al pasar de un virus a otro. Sin embargo, esto no ocurre, tal y como mostramos en la figura 4.4, en donde hemos superpuesto cuatro mapas de correlación genética correspondientes a cuatro genes *env* pertenecientes cada uno a un VIH-1 distinto.

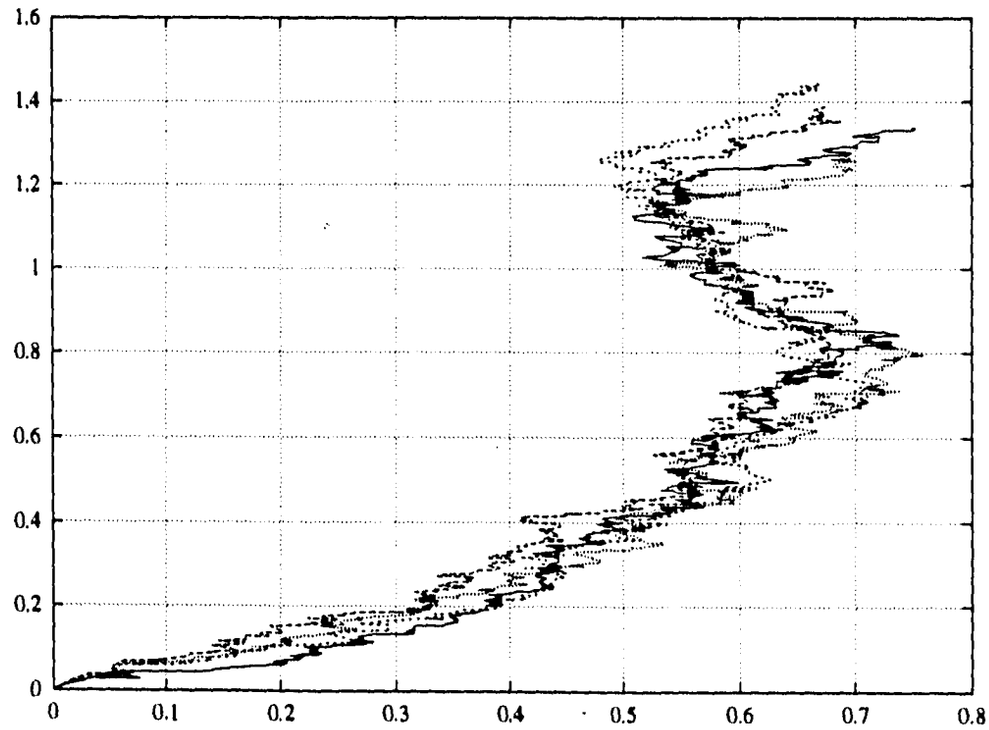


Figura 4.4

Superposición de cuatro mapas de correlación genética correspondientes cada uno a un virus de SIDA tipo 1 diferentes. Notemos la gran similitud de estructura que presentan estos cuatro mapas, en los cuales se ha tomado $d_{max} = 500$.

Las secuencias genéticas de estos cuatro genes fueron tomadas del GENE BANK 94 de manera arbitraria (esto es, no fueron escogidos), y vemos que en los cuatro casos, el mapa de correlación genética presenta la misma estructura o forma general. Los cuatro mapas no son exactamente idénticos, sino que presentan "fluctuaciones", lo cual refleja el hecho de que han ocurrido mutaciones puntuales en las secuencias genéticas de estos genes. No obstante, el que la estructura general del mapa genético sea la misma para los cuatro genes, refleja que, pese a las mutaciones, hay una estructura general conservada en las secuencias genéticas que conforman a los genes *env* de los diferentes virus.

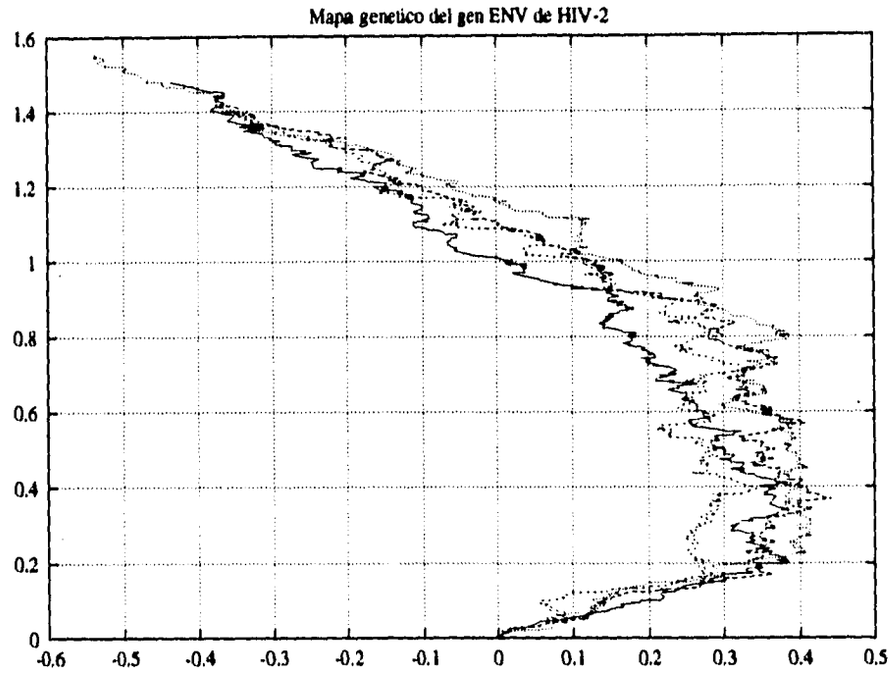
Y bueno, ya que teníamos el programa que generaba los mapas de correlación genética, lo más fácil era aplicarlo a diferentes secuencias genéticas. En las figuras 4.5(a) y 4.5(b) mostramos los mapas correspondientes al gen *env* del VIH–2 y al gen *env* del virus que causa el SIDA en el mono verde africano (SIV). Vemos de estas figuras que a pesar de las mutaciones puntuales que hay en los genes también existe una estructura global que no cambia, tanto en el gen *env* del VIH–2 como en el gen *env* del SIV. Pero si comparamos las figuras 4.4 y 4.5 nos damos cuenta de algo de la mayor importancia, a saber, que el mapa genético correspondiente al gen *env* del VIH–1 es sustancialmente diferente al mapa del gen *env* del VIH–2, y estos dos son diferentes del mapa correspondiente al gen *env* del SIV. En otras palabras, lo que estos mapas están reflejando es que las estructuras conservadas, sean las que sean, en los genes *env* pertenecientes ya sea al VIH–1, o al VIH–2, o al SIV, son propias y específicas de cada uno de los tres tipos de virus. Y decimos "sean las que sean" porque en el momento de escribir esta Tesis, no tenemos muy en claro cómo encontrar dichas estructuras conservadas a partir de lo mapas genéticos. Estos mapas nos indícan claramente que tales estructuras existen, pero ¿cuáles son? no lo sabemos aún.

En el apéndice de este trabajo damos un "atlas" de mapas de correlación genética del VIH, es decir, mostramos los mapas correspondientes a los genes *env*, *gag* y *pol* tanto de VIH–1 como de VIH–2 y de SIV, así como los correspondientes a otras secuencias que se explican en el apéndice. Por el momento mostramos aquí en la figura 4.6 otros dos mapas de correlación genética: el mapa de la figura 4.6(a) corresponde al genoma completo del VIH–1 y el de la figura 4.6(b) corresponde a la secuencia genética que contiene al grupo de genes que codifican para el cúmulo β de la hemoglobina en el ser humano (ver la figura 2.1(a))†. Otra vez volvemos a ver una diferencia estructural entre ambos mapas, ya que si comparamos el mapa de la figura 4.6(b) con el de la figura 4.2(b) vemos que en la secuencia intergénica de la β–globina existen correlaciones de corto y largo alcance que no están presentes en el genoma del virus del SIDA.

---

† Recordemos que esta secuencia genética está constituída en su mayor parte por secuencias intergénicas:menos del 2% de la secuencia es codificadora. Por lo tanto, podemos referirnos a ella como "la secuencia intergénica del cúmulo β de la hemoglobina".

(a)



(b)

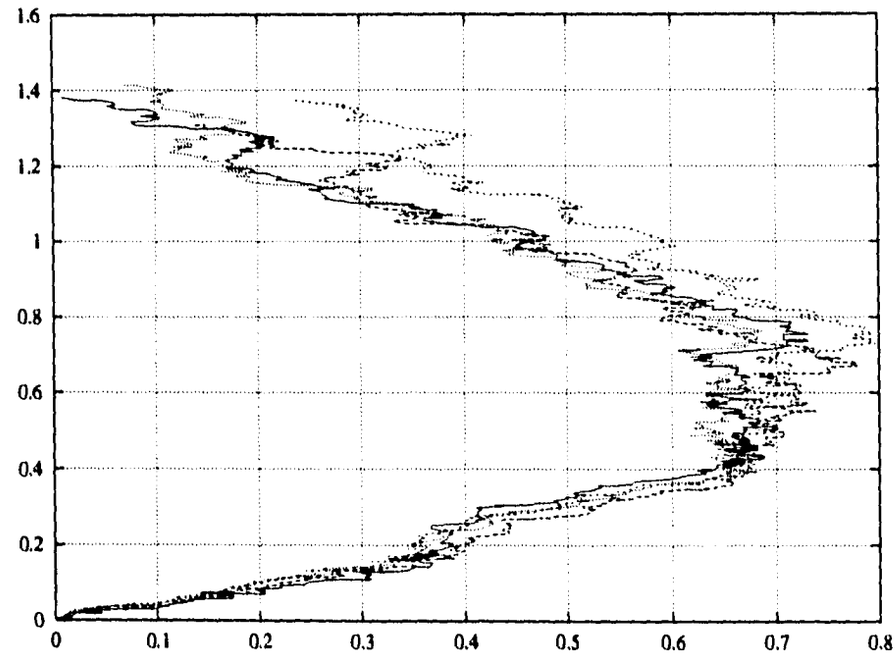
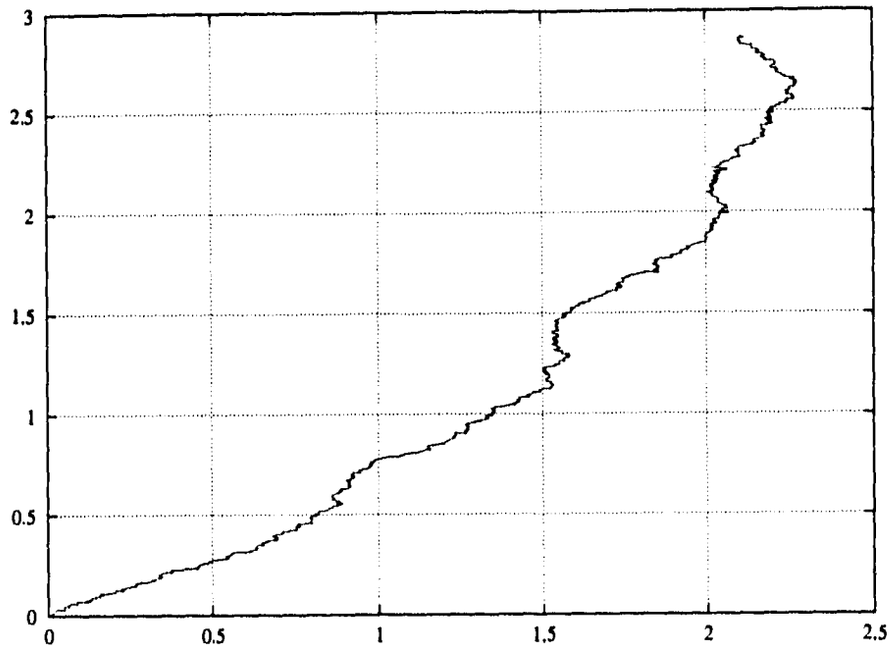


Figura 4.5

(a) Superposición de cuatro mapas de correlación genética correspondientes cada uno a la secuencia genética del gen env de cuatro VIH-2 diferentes. (b) Superposición de cuatro mapas correspondientes al gen env de cuatro SIV diferentes. En todos los mapas se tomó $d_{max} = 500$. *Nótese la diferencia de escalas en (a) y (b).*

(a)



(b)

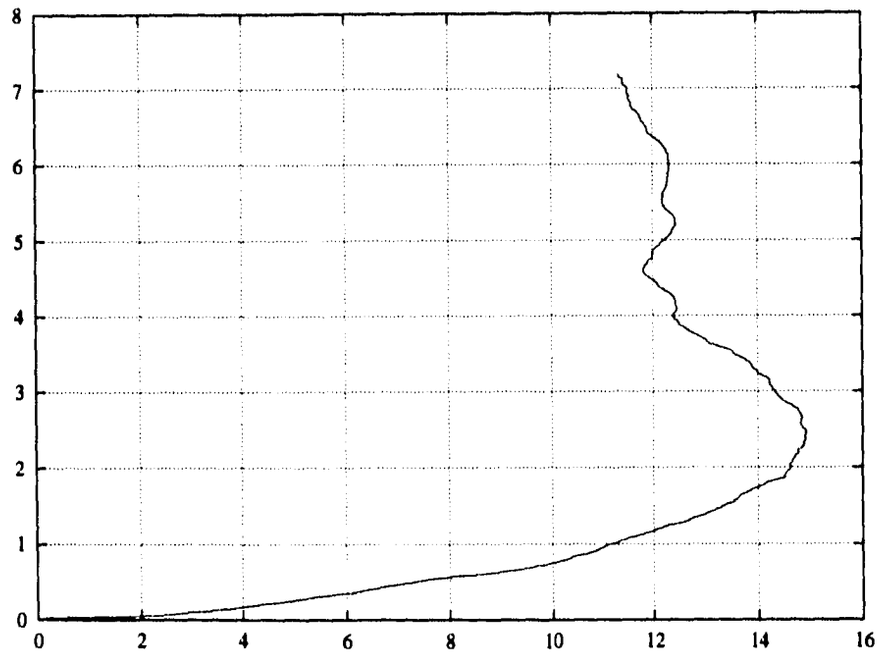


Figura 4.6

(a) Mapa de correlación genética correspondiente al genoma completo de un VIH-1, con $d_{max} = 500$. *(b) Mapa de correlación genética correspondiente a la secuencia intergénica del cúmulo* β *de la hemoglobina del sel humano con* $d_{max} = 15000$.

Origen de la parte imaginaria de los eigenvalores.

Terminaremos este capítulo dando una explicación de por qué para ciertas distancias los eigenvalores de la matriz de correlación **C** tienen parte imaginaria. Para esto, notemos que si la matriz **C** fuera simétrica en todos los casos entonces sus eigenvalores serían siempre reales. Esto sugiere que la parte imaginaria de los eigenvalores puede surgir de la "componente" antisimétrica de la matriz. En el caso de una matriz arbitraria lo anterior no necesariamente ocurre, ya que si bien es cierto que la simetría de la matriz implica que sus eigenvalores sean reales, la implicación en sentido contrario en general no ocurre, es decir, no siempre que los eigenvalores son reales la matriz es simétrica. Por lo tanto, la propiedad que estamos demandando de nuestra matriz de correlación **C**, de que la parte imaginaria de sus eigenvalores surja de la componente antisimétrica de dicha matriz, es bastante fuerte.

Expresemos pues a la matriz **C** en términos de sus componentes simétrica (**S**) y antisimétrica (**A**), escribiendo a **C** como

$$\mathbf{C}(d) = \mathbf{S}(d) + \mathbf{A}(d)$$

donde

$$S_{\alpha\beta}(d) = \frac{1}{2}\big(C_{\alpha\beta}(d) + C_{\beta\alpha}(d)\big)$$

$$A_{\alpha\beta}(d) = \frac{1}{2}\big(C_{\alpha\beta}(d) - C_{\beta\alpha}(d)\big)$$

Es claro que con esta descomposición de **C** se cumple que

$$Tr[\mathbf{C}(d)] = Tr[\mathbf{S}(d)]$$

Ahora bien, como la matriz **A** es antisimétrica, sus eigenvalores serán imaginarios puros. Si llamamos λ_A a los eigenvalores de **A**, podemos escribirlos como

$$\lambda_{A1} = ic$$

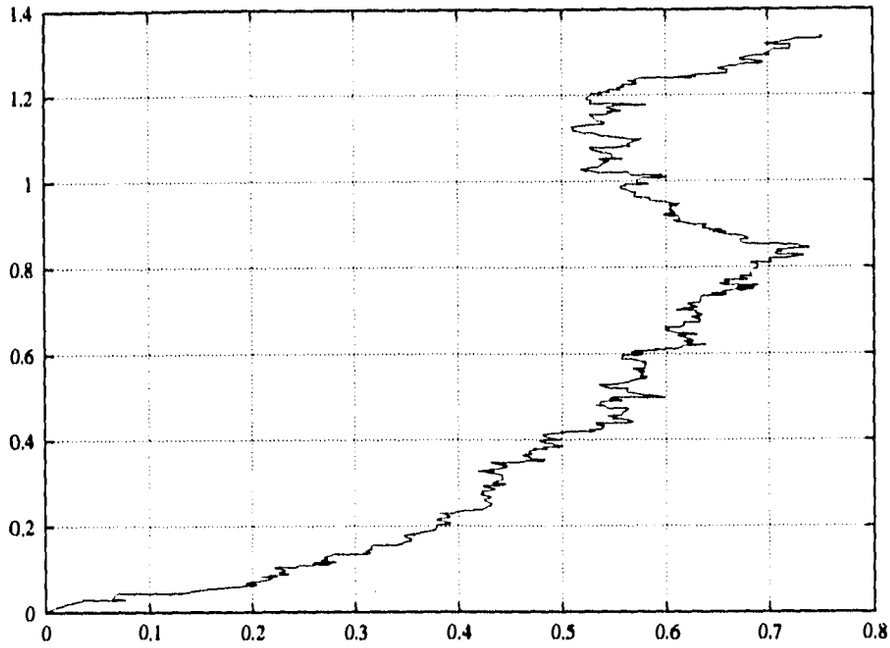
$$\lambda_{A2} = -ic$$

siendo c un número real positivo. Construímos ahora el mapa de correlación genética sumando los vectores $\vec{w}(d)$ definidos como

$$\vec{w}(d) = \big(Tr[\mathbf{C}(d)]\ ,\ c\big)$$

Con esto, nuestro nuevo vector de eigenvalores $\vec{w}(d)$ tiene como primer componente a la traza de la matriz de correlación **C**, y como segunda componente estamos poniendo a la magnitud los eigenvalores de la parte antisimétrica de **C**.

(a)



(b)

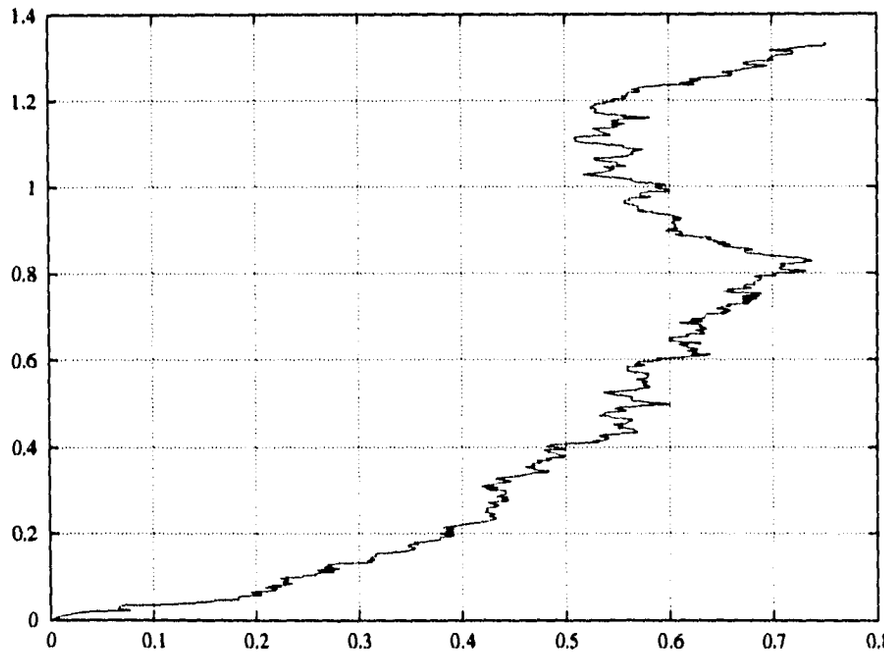


Figura 4.7

(a) Mapa de correlación genética correspondiente a la secuencia genética del gen env de un VIH-1 calculado con los vectores de eigenvalores $\vec{v}(d)$. (b) Mapa correspondiente a la misma secuencia que en (a), pero calculado con los vectores de eigenvalores $\vec{w}(d)$ relacionados con la componente antisimétrica de la matriz de correlación.

En la figura 4.7(a) mostramos el mapa de correlación genética construído con los vectores $\vec{v}(d)$ como antes, correspondiente a la secuencia del gen *env* de un VIH–1, y el la figura 4.7(b) mostramos el mapa de correlación genética correspondiente a la misma secuencia, pero construído ahora con los vectores $\vec{w}(d)$. Podemos ver que estos dos mapas son escencialmente idénticos, lo que confirma nuestra suposición de que la parte imaginaria de los eigenvalores de $\mathbf{C}(d)$ proviene exclusivamente de su componente antisimétrica.

Por lo tanto, en el caso cuando los eigenvalores de $\mathbf{C}(d)$ son reales, podemos afirmar que

$$\mathbf{A}(d) \simeq 0$$

o bien

$$C_{\alpha\beta}(d) - C_{\beta\alpha}(d) \simeq 0 \qquad \alpha, \beta = 1, 2, 3, 4$$

y si utilizamos la definición (4.2) del índice de correlación, lo anterior se transforma en

$$N_{\alpha\beta}(d) \simeq N_{\beta\alpha}(d)$$

es decir, cuando a la distancia d los eigenvalores de $\mathbf{C}$ son reales, hay el mismo número de parejas $\alpha\beta$ que de parejas $\beta\alpha$. De lo anterior podemos deducir que la parte compleja de los eigenvalores contiene a las correlaciones no simétricas, es decir, aquellas para las cuales $N_{\alpha\beta} \neq N_{\beta\alpha}$. Esto es un resultado de la mayor importancia, ya que nuestro análisis está reflejando que las cadenas de ADN tienen una dirección de lectura, que no es lo mismo leer el mensaje en una dirección (del extremo 3' al extremo 5') que en la dirección opuesta (del extremo 5' al 3'), cosa que ya sabíamos por princípios biológicos, pero que nuestras técnicas de análisis están mostrando, a través de la parte antisimétrica de la matriz de correlación, utilizando para ello solamente a la secuencia de bases del ADN.

Conclusiones

Aquello a lo que doy forma a la luz del día es solamente el uno por ciento de lo que he visto en la obscuridad.
M. C. Escher.

Hemos llegado al final de nuestro trabajo y tenemos que concluir, no porque hallamos agotado el tema, ni mucho menos, sino porque en algún momento teníamos que "cortar" esta tesis. Sin embargo, aún queda mucho por hacer.

Por una parte, todavía no sabemos como pasar de las estructuras conservadas en los Mapas de Correlación Genética a las estructuras conservadas en las secuencias genéticas reales. Estos mapas nos señalan muy claramente que tales estructuras conservadas existen dentro de las secuencias genéticas, pero aún no sabemos cuáles son. No obstante, estos mapas nos han servido para encontrar "parentesco" entre diferentes organismos. Por ejemplo, si observamos con detenimiento los Mapas de Correlación Genética que aparecen en el Apéndice, veremos que los mapas correspondientes a los genes del HIV-2 son muy similares (casi idénticos excepto, por un factor de escala) a los mapas correspondientes a los genes del SIV, pero por otro lado, tienen una estructura muy diferente que la de los mapas correspondientes al HIV-1. Podemos concluir, a partir de lo anterior, que el HIV-2 y el SIV están "emparentados" (en el sentido de que el HIV-2 pudo provenir del SIV a través de mutaciones), más no así con el HIV-1 y el HIV-2. Lo mismo podemos decir de las secuencias intergénicas pertenecientes al cúmulo β de la hemoglobina en el ser humano y en el conejo. Es decir, nuestros mapas nos revelan estructuras y parentescos que no podían hacerlo los métodos tradicionales de la Biología, tales como los índices de variabilidad, o los índices de homología.

Más aún, hasta ahora nuestros Mapas de Correlación Genética nos dan estructuras y parentescos en y entre las secuencias genéticas "a ojo", es decir, sólo gráficamente. Sin embargo, debemos de ser más formales e idear una forma analítica que nos diga que "tan cerca" o que "tan lejos" se encuentran dos mapas uno de otro. Huelga decir que aún no hemos ideado dicha forma analítica. Lo que si podemos hacer es, dada una secuencia genética, decir si en ella hay correlaciones de largo alcance o no, o si esta secuencia pertenece al gen *env* de un virus de SIDA tipo 1 o al gen *pol* de un virus de SIDA tipo 2 (por

ejemplo), simplemente por inspección del Mapa de Correlación Genética correspondiente a dicha secuencia. Nuestros mapas también sirven como un "test" externo para checar la consistencia de los algoritmos genéticos de mutación, los cuales tratan de obtener una secuencia genética de algún organismo real, a partir de una secuencia generada al azar, produciendo mutaciones con probabilidades bien determinadas, en la secuencia aleatoria.

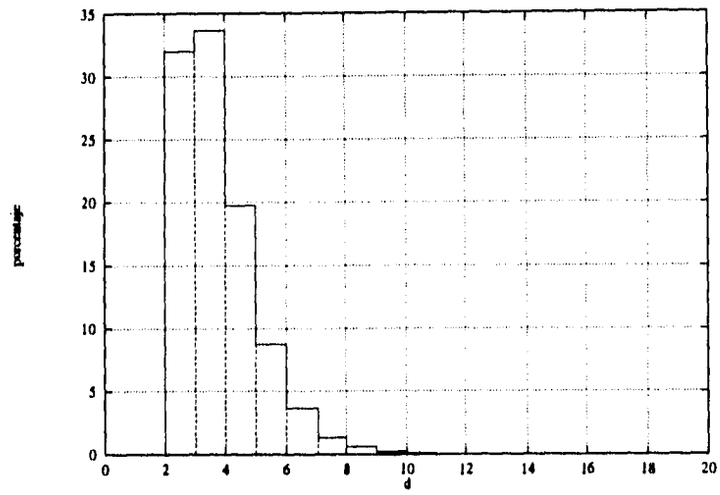
Por otro lado, en este trabajo hemos llegado a una conclusión de la mayor importancia dentro de la Biología Molecular: la información genética contenida en los ácidos nucleicos se lee por codones, es decir, de tres bases en tres bases, porque los mínimos consecutivos de la energía de amarre entre parejas de bases se encuentran espaciados a lo largo de la secuencia genética, en promedio, cada tres bases, y esto es independiente de los valores numéricos de las energías de amarre entre parejas de bases, así como del número de bases con las cuales se construyan las moléculas portadoras de la información genética (mientras sean más de dos bases). Como hemos visto, el que la distancia promedio entre mínimos consecutivos de energía a lo largo de la secuencia sea 3, sólo depende de que exista un princípio de complementariedad entre parejas de bases y de que exista una relación de "...mayor que..." entre los valores numéricos de tales energías. Hasta donde sabemos, lo anterior es la primera justificación fundamentada de la naturaleza del Código Genético.

El trabajo que hicimos sobre lo del periodo 3 fue en su mayoría numérico. Sin embargo, creemos que los resultados que obtuvimos pueden demostrarse en forma analítica formando la función de partición del sistema (la secuencia genética es nuestro sistema) y demostrando que las configuraciones que más contribuyen a la función de partición son aquellas en las cuales los mínimos de energía están separados cada tres bases, pero hasta el momento no hemos podido demostrarlo de esta forma.

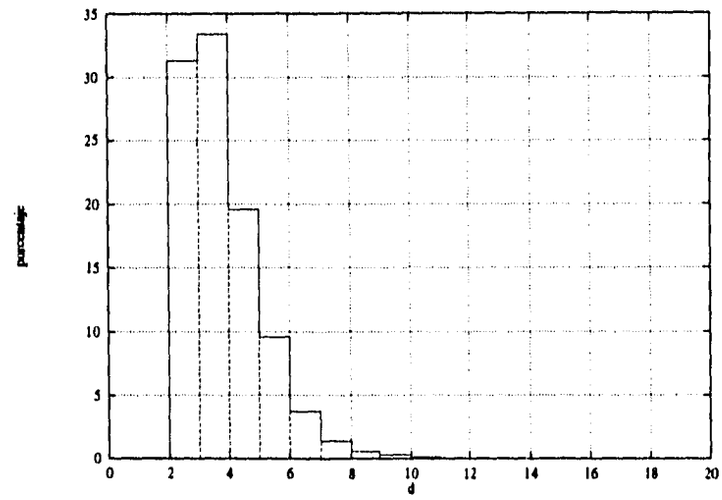
Todavía resta mucho por hacer. Las leyes de la Física y de la Química nos dicen como *interactúa* la materia. Pero la materia no solamente interactúa, sino que también se *organiza*, y lo que no tenemos son las leyes de organización de la materia, indispensables para el entendimiento de los organismos vivos. El autor espera honestamente que este trabajo contrubuya, aunque sea sólo un poco, en la búsqueda de estas leyes de organización.

Apéndice

(a)

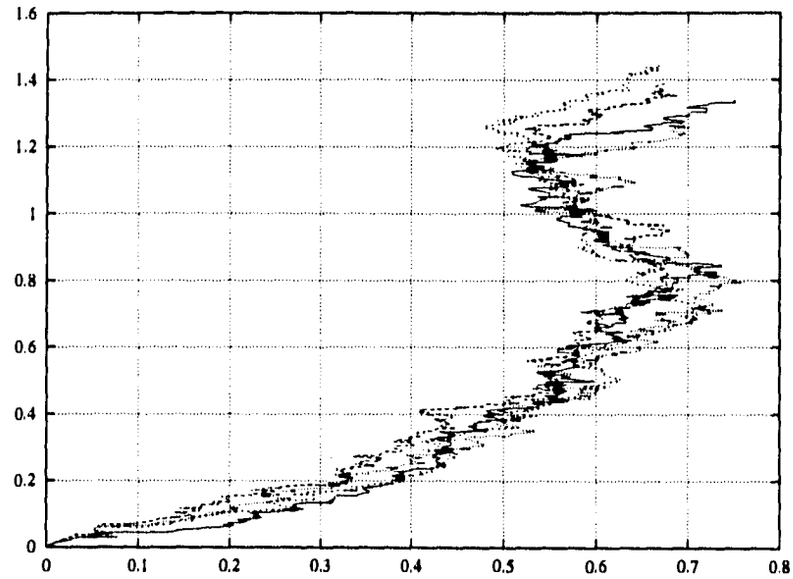


(b)

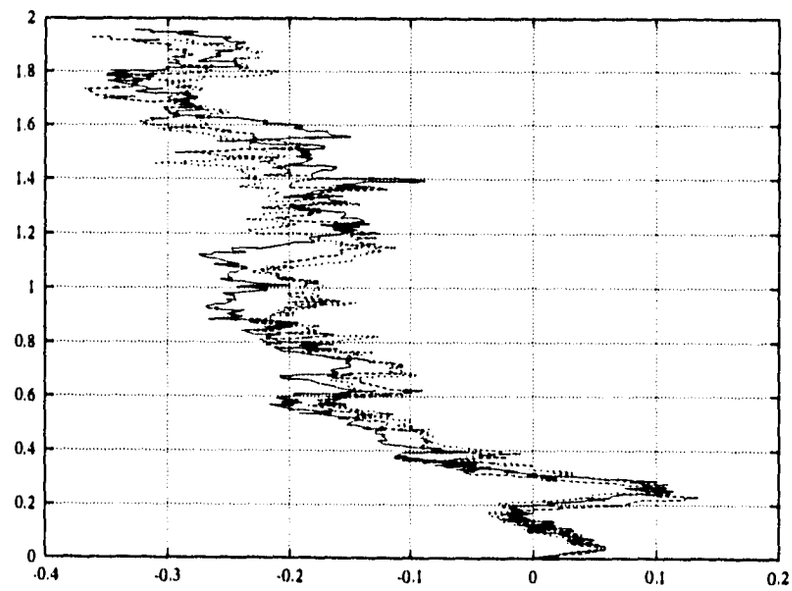


Distribucion de la distancia d entre mínimos de energía consecutivos a lo largo de la secuencia intergénica del cúmulo β de la hemoglobina (a) en el ser humano y (b) en el conejo. Vemos que para secuencias genéticas reales no solamente vale 3 el promedio de la distancia, sino que también vale 3 la moda. El histograma está dado en unidades de 100%. Las energías utilizadas fueron las reportadas por Breslauer para ΔH del ARN.

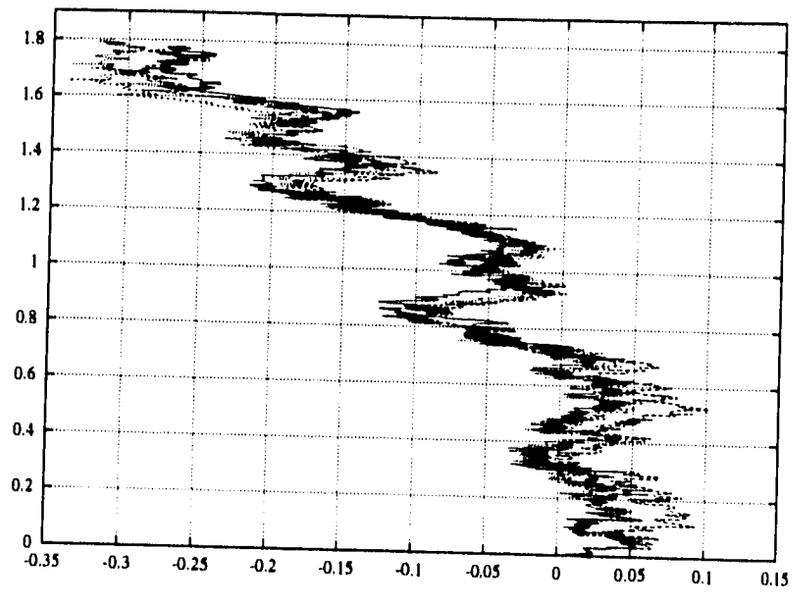
Mapas de Correlación genética



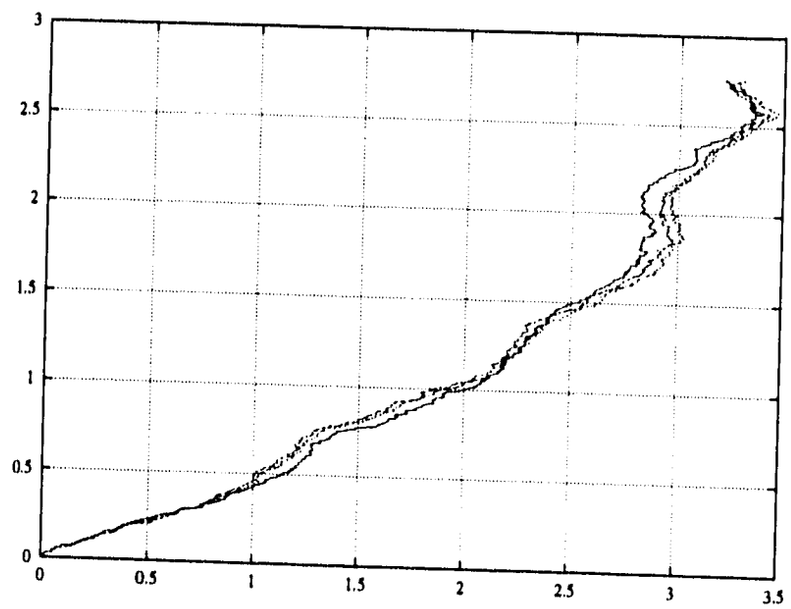
Gen *env* de HIV-1. Distancia máxima $d_{max} = 500$.



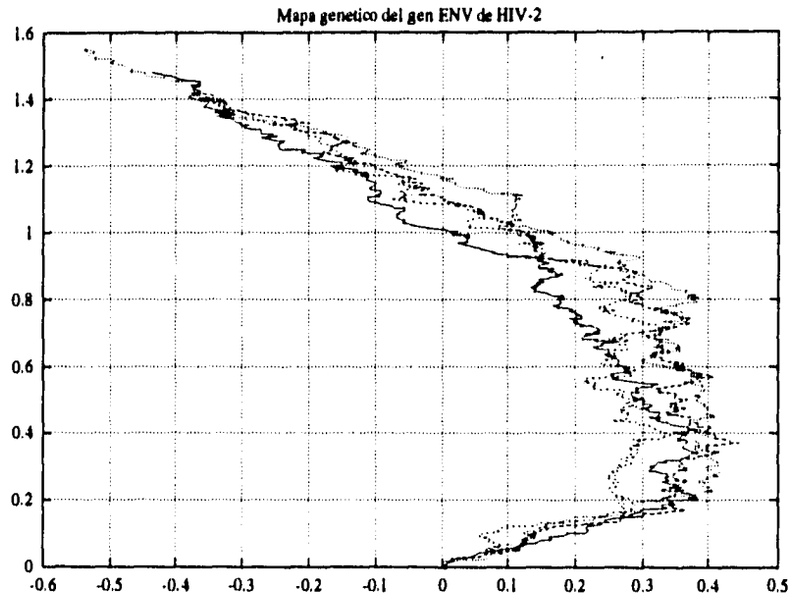
Gen *gag* de HIV-1. $d_{max} = 500$.



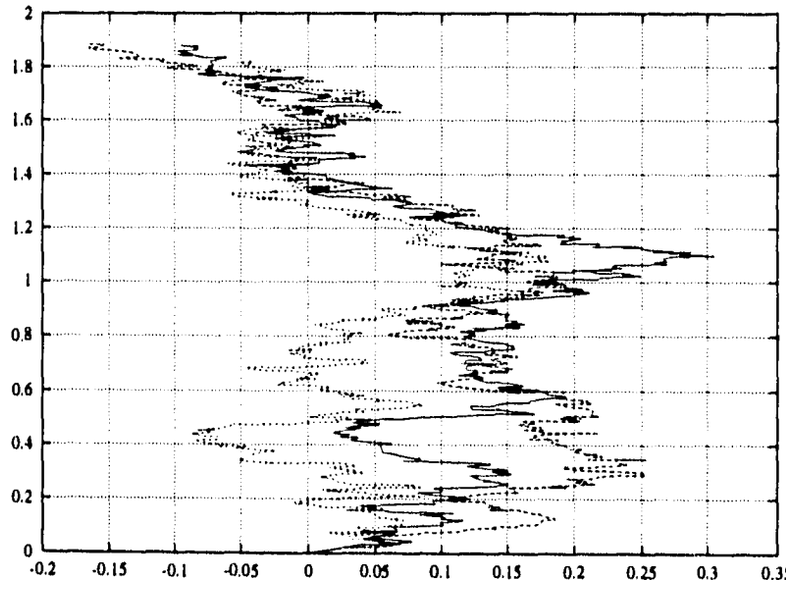
Gen *pol* de HIV-1. $d_{max} = 500$.



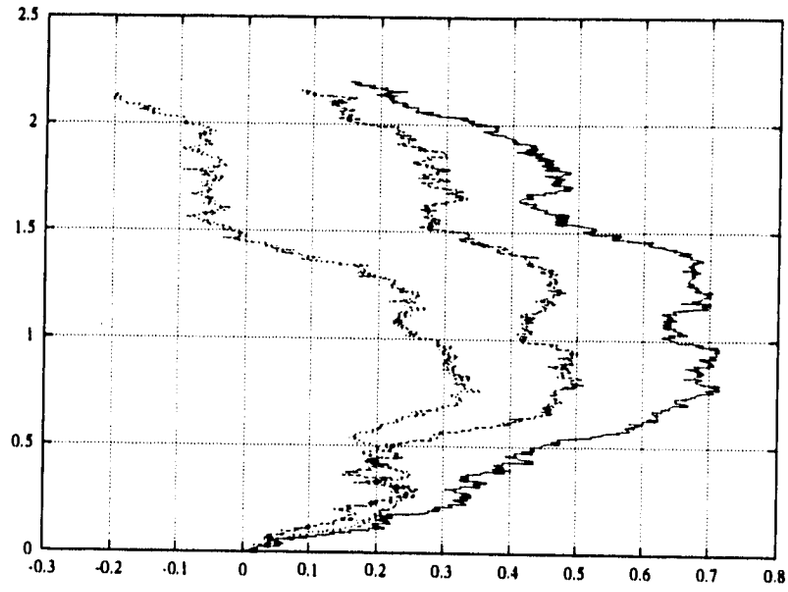
Genoma completo de HIV-1. $d_{max} = 2000$.



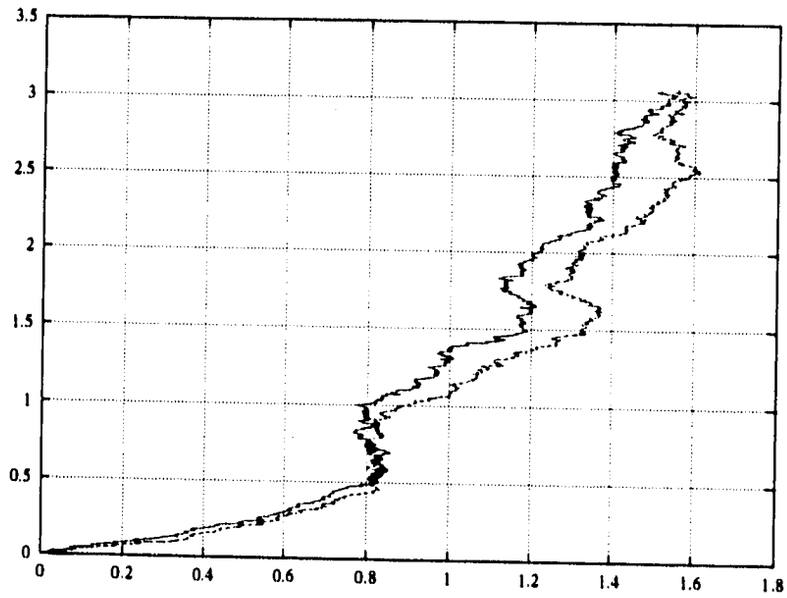
Gen *env* de HIV–2. $d_{max} = 500$.



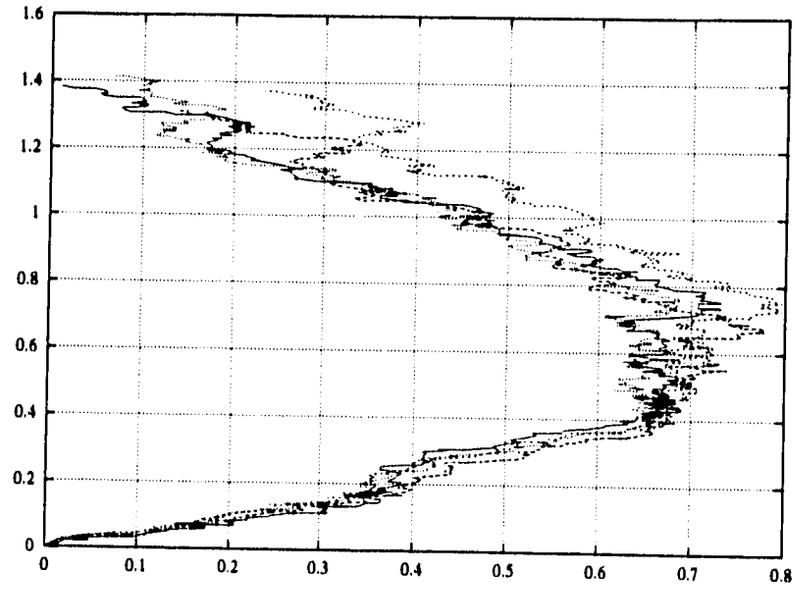
Gen *gag* de HIV–2. $d_{max} = 500$.



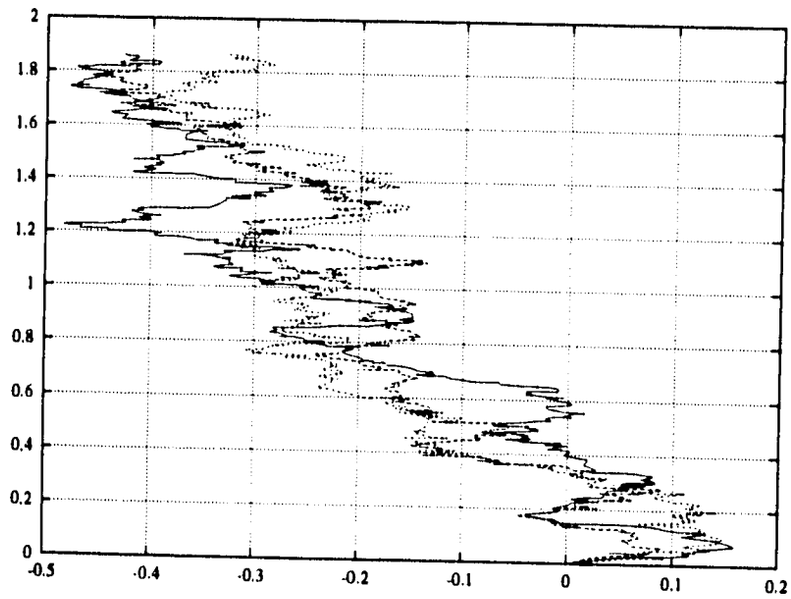
Gen *pol* de HIV–2. $d_{max} = 500$.



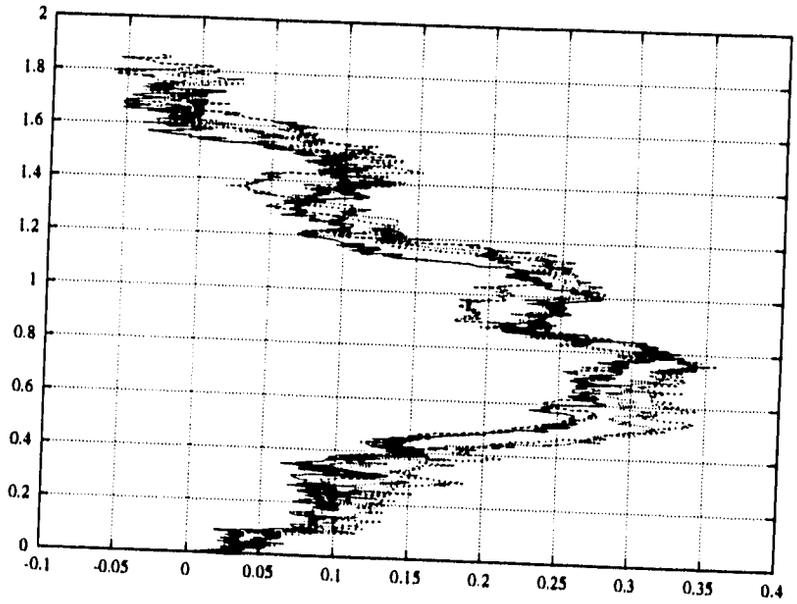
Genoma completo de HIV–2. $d_{max} = 2000$.



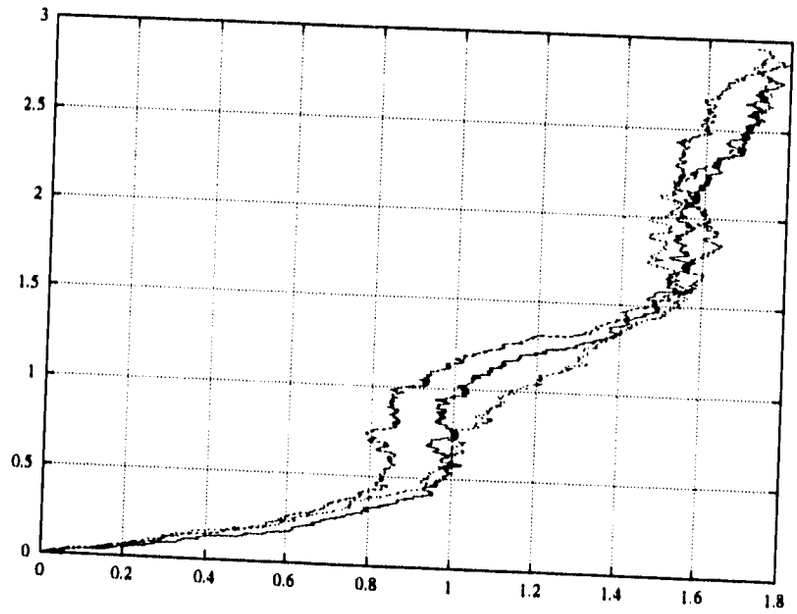
Gen *env* de SIV. $d_{max} = 500$.



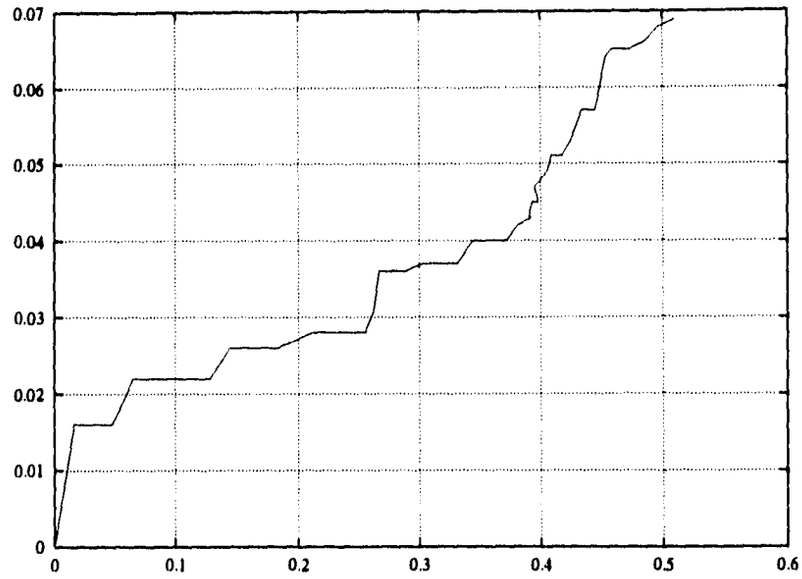
Gen *gag* de SIV. $d_{max} = 500$.



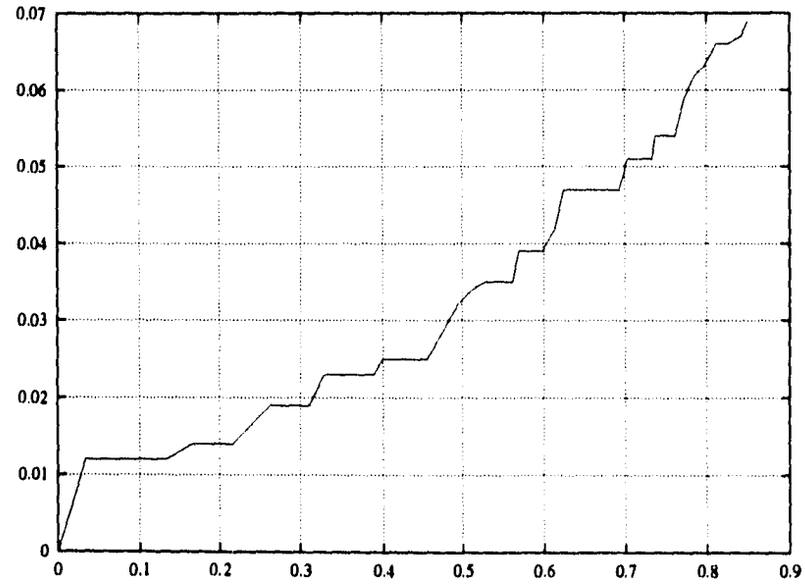
Gen *pol* de SIV. $d_{max} = 500$.



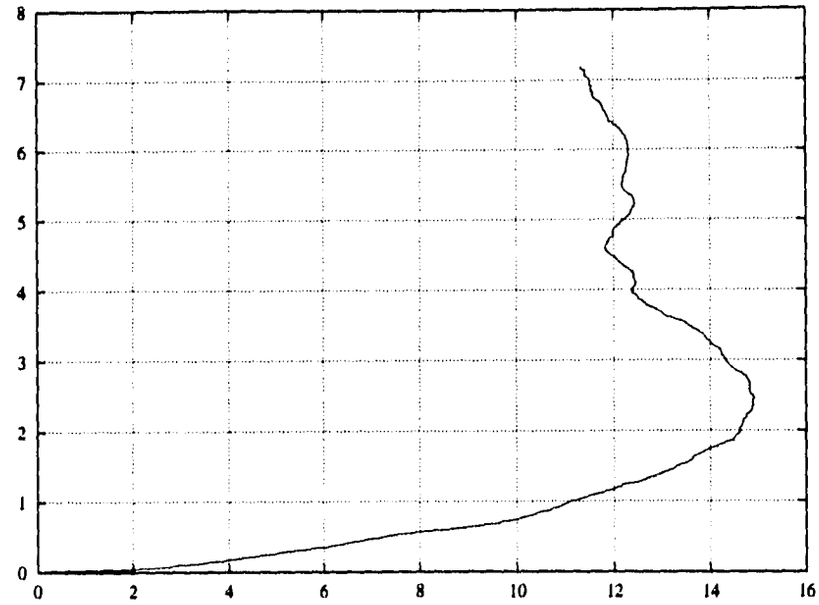
Genoma completo de SIV. $d_{max} = 2000$.



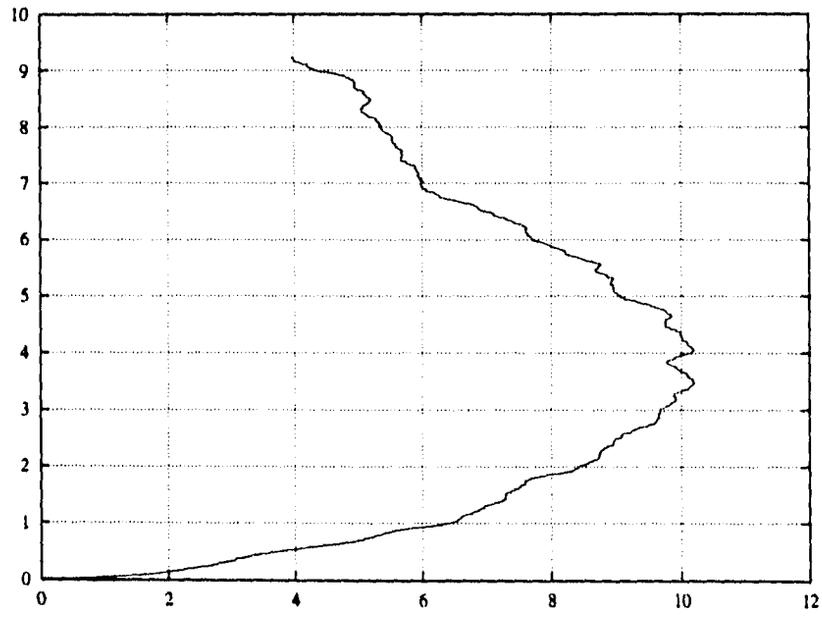
Secuencia intergénica del cúmulo β de la hemoglobina en el ser humano. $d_{max} = 50$.



Secuencia intergénica del cúmulo β de la hemoglobina en el conejo. $d_{max} = 50$.



Secuencia intergénica de la β-globina en el ser humano. $d_{max} = 11000$.



Secuencia intergénica de la β-globina en el conejo. $d_{max} = 11000$.

Referencias

[1] Watson, J. D., y Crick, F. H., 1953a. *Molecular structure of nucleic acid. A structure for deoxysibose acid.* Nature, 171: 737–738.

[2] Watson, J. D., y Crick, F. H. C., 1953b. *Genetic implications of the structure of deoxyribonucleic acid.* Nature, 171: 964–967.

[3] Niremberg, M., 1968. *The genetic code.* En Nobel Lectures: Physiology or Medicine (1963–1970), pp 372–395.

[4] Crick, F. H. C., Barnet, L., Bernner, S., y Waltts-Tobin, R. J., 1961. *General nature of the genetic code for proteins.* Nature, 192: 1227–1232.

[5] Cochet, M., Gannon, F., Hen, R., Maroteaux, L., Perrin, F., y Chambon, P., 1979. *Organization and secuence sudies of the 17-piece chicken conalbumin gene.* Nature, 282: 567–574.

[6] Cech, t. R., y Bass, B. L., 1986. *Biological catalisis by RNA.* Ann. Rev. Biochem: 55:599.

[7] Temin. H., 1976. *The DNA provirus hypotesis: the establyshment and implications of RNA-directed DNA synthesis.* Science, 192: 1075–1080.

[8] Baltimore, D., 1976. *Virus, polymerases and cancer.* Sience, 192: 632–636.

[9] Greene, W. C., 1993. *SIDA y el sistema inmunitario.* Investigación y Ciencia, 206: 58–66.

[10] Reyes-Terán, G., y Ponce de León, S., 1994. *SIDA: los laberintos de la infección.* Ciencias, 33: 31-42.

[11] Janeway, C. A. Jr., 1993. *Reconocimiento inmunitario de cuerpos extraños.* Investigación y Ciencia, 206: 26–33.

[12] Breslauer, K. J., Ronald Frank, Blöcker, H., y Marry, L. A., 1986. *Predicting DNA duplex stability fron the base sequence.* Proc. Natl. Acad. Sci. USA, 83: 3746–3750.

[13] Magnasco, M. O., 1994. *Molecular combustion motors.* Physical Review Letters, 72: 2656–2659.

[14] Millonas, M. M., y Dykman, M. I., 1994. *Transport and current reversal in tochastically driven ratches.* Physics Letters A, 185: 65–69.

[15] Manfred Eigen, 1993. *Viral Quasispecies.* Scientific American, 269: 32–39.

[16] Kaneko, K., Wentran Li, Marr, T. G., 1994. *Understanding long-range correlations in DNA sequences.* Physica D, febrero 1994.