UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS
DIVISION DE ESTUDIOS DE POSGRADO

UN ESQUEMA DE AUTOMATA CELULAR COMO MODELO MATEMATICO DE LA EVOLUCION DE LOS ACIDOS NUCLEICOS

TESIS

QUE PARA OBTENER EL GRADO DE

DOCTOR EN CIENCIAS (MATEMATICAS)

PRESENTA

PEDRO EDUARDO MIRAMONTES VIDAL

MÉXICO, D.F. 1992

INTRODUCCION

Hasta hace aproximadamente un par de décadas, la mayoría de los científicos dedicados al modelaje matemático de fenómenos naturales compartían la idea de que la complejidad matemática del modelo tenía que ir aparejada a la complejidad del fenómeno en cuestión; es decir, que el modelo matemático de un fenómeno sumamente complicado debería ser matemáticamente muy complejo. Esta manera de pensar dejaba de lado una gran cantidad de fenómenos interesantes de la naturaleza, debido a que el aparato matemático resultaba intratable.

Esta cosmovisión no es ajena al enfoque reduccionista de las ciencias; en efecto, si se parte de la premisa de que la labor esencial de la ciencia es la de encontrar leyes "fundamentales" y, a partir de ellas, estudiar, interpretar y entender la naturaleza, la única manera posible de hacerlo es a través de un proceso inverso al de la búsqueda de las leyes; a través de un proceso constructivista. Según este esquema, hace falta, en última instancia, resolver lo concerniente a la física de partículas elementales de cualquier fenómeno y después, gradualmente, si se quiere, pasar al estudio de los niveles superiores de organización de la materia. El enfoque reduccionista ha tenido éxito en la Física. Sin embargo, poca gente piensa seriamente en explicar la conducta de un organismo, comunidad o sociedad (vamos, ni de una simple célula) a partir de las ecuaciones de las particulas elementales.

En la década de los setentas, a partir del trabajo de May [1] y, de manera especialmente meritoria, del de Anderson [2] este punto de vista comenzó a cambiar gradualmente; Se empezó a aceptar que los grandes agregados de partículas (partículas elementales, átomos, células, etc.) podrían tener comportamientos propios, inherentes al conjunto y que no se pueden deducir a partir de leyes "fundamentales". Empezaron a acuñarse nuevos

términos tales como el de "ruptura de simetría" [2] y se emprendió la búsqueda de las leyes fundamentales para cada nivel de complejidad o de organización de la materia. Adicionalmente, se comenzó a aceptar la idea de que algunos esquemas matemáticos relativamente simples podían dar lugar a dinámicas complicadas. La propuesta de May, originalmente pensada como un modelo muy simple para la dinámica poblaciones, es un modelo continuo en el rango de la variable de estado pero discreto en el tiempo.

Recientemente [3,4], se han empezado a estudiar fenómenos en los cuales el número de variables de estado es demasiado pequeño como para poder estudiarse estadísticamente o como para poder suponer un continuo, pero en cambio, es suficientemente grande como para poder intentar la deducción o solución de ecuaciones dinámicas. Entre estos fenómenos, se hallan multitud de problemas de optimización, adaptación, aprendizaje, evolución, etc. a los cuales se ha dado en llamar "sistemas complejos" [4]. Su surgimiento viene aunado a un florecimiento reciente en las matemáticas discretas; una muestra pequeña de estos métodos incluye las redes neuronales, los algoritmos genéticos, los autómatas celulares y muchos más.

Aunque la palabra "complejidad" tiene muchas acepciones diferentes, se acepta que un sistema es "complejo" si tiene un gran número de grados de libertad y un gran número de soluciones dinámicas diferentes, tanto en lo cualitativo como en lo cuantitativo. Un factor importante en la generación de muchas soluciones diferentes a un problema lo constituye la presencia de dinámicas en conflicto; en otras palabras, la ocurrencia antagónica de interacciones o correlaciones de diferente rango y naturaleza actuando simultáneamente sobre un conjunto discreto de elementos [5]. Cada uno de estos elementos puede ser, a su vez, un sistema complejo, pero si se supone que cada uno de ellos tiene solamente un conjunto pequeño de rasgos característicos relevantes, entonces pueden ser estudiados como algo sim-

ple. Podemos mencionar como ejemplos de sistemas complejos a los vidrios de espín [6] en donde los elementos discretos son átomos o moléculas, al problema de la Morfogénesis [7] (células como elementos discretos, contacto entre ellas como interacciones de corto alcance y difusión de morfógenos como interacciones de largo alcance), sistemas sociales (individuos como elementos, con contactos personales o familiares inmediatos e influencias de largo alcance debidas a los medios de comunicación masiva).

En la dinámica del genoma pueden presentarse conflictos tanto de largo como de corto alcance; los unos debido a interacciones entre nucleótidos vecinos y los otros como consecuencia de las interacciones de largo alcance asociadas a la traducción del DNA en proteínas funcionales[1].

Los autómatas celulares ocuparán un sitio importante en este trabajo; se caracterizan como sistemas dinámicos con interacciones de corto alcance (locales) en los cuales se encuentran discretizados el tiempo, las variables de estado y las variables espaciales. Se estudiarán, además, mecanismos de optimización global que pueden entenderse como restricciones de largo alcance.

En el capítulo I de este trabajo se da una introducción a la biología de los ácidos nucléicos, en particular del ADN. La intención de este capítulo es la de familiarizar al lector no versado en biología con la fenomenología de la genética molecular para lograr establecer un lenguaje común y poder así, introducir un modelo matemático de la evolución filogenética de la molécula de ADN.

En el capítulo II, se presenta una introducción a los sistemas dinámicos llamados autómatas celulares que parecen ser

---

[1] Buena parte de la actividad de las proteínas depende de su estructura tridimensional. En esta, aminoácidos que pueden provenir de sitios muy apartados el la secuencia del ADN, pueden estar colocados en posiciones espacialmente muy cercanas.

una herramienta matemática adecuada para el modelaje de la evolución de moléculas grandes en general porque tienen la característica de ser discretos tanto en el tiempo como en el espacio y en el rango de la variable.

En el capítulo III, haciendo uso de los citados autómatas celulares, se propone formalmente un modelo de desarrollo filogenético de la molécula de ADN. Se enfoca esta propuesta desde un punto de vista fisicalista y se exploran ciertas reguaridades en la distribución local de energía de la molécula de ADN para proponer reglas de evolución temporal. Es necesario aclarar que de aquí en adelante, el término "evolución" puede tener dos acepciones diferentes; la primera referente a la evolución biológica de los seres vivos y la segunda relativa a la evolución en el tiempo del sistema dinámico. Cuando el contexto sea suficientemente claro para que no haya confusión, no se hará la distinción explícita.

En el capítulo IV, se proponen las condiciones para que un conjunto de mutaciones sea capaz de aumentar la adecuación (la capacidad de aumentar la tasa reproductiva) de un organismo y de esta manera se integre al genoma. La prueba se basa en un proceso de optimización discreta, se ha elegido para este propósito el método de los llamados *algoritmos genéticos*.

El trabajo concluye con un breve capítulo V de resultados y conclusiones.

Es extremadamente fácil caer en el lenguaje que refleja el estilo dominante de la Biología, estilo acorde con los esquemas políticos dominantes y que prevalece tanto en la mentalidad como en la literatura biológica [8]. Este lenguaje nos ha acostumbrado a visualizar (incluso a concebir) la evolución biológica como un proceso de costo-ganancia, o bien, de inversión-beneficio. Cualquier frase en el presente trabajo que cause esta impresión, es descuido y no convicción del autor.

Este trabajo es parte de un proyecto de investigación sobre la estructura y la función del ADN, dirigido por el Dr. Germinal Cocho y en el cual participan también los Maestros Luis Medrano y Antonio Lazcano y los Doctores José Luis Rius, Gustavo Martínez y Radmila Bulajich. El autor desea agradecer la inapreciable dirección del Dr. Cocho así como la cuidadosa lectura y comentarios del M. en C. Medrano. Concha Ruiz le dedicó interminables horas a la revisión de estilo de manuscrito, por esta y por otras muchas razones, el autor le dedica, con amor y gratitud, el trabajo entero.

CAPITULO I

FENOMENOLOGIA

La molécula de ADN es un polímero lineal (Figura 1) construido como una sucesión de cuatro monómeros distintos llamados nucleótidos. Cada monómero tiene una parte constante (moléculas de fosfato y desoxirribosa) y una parte variable que puede ser cualquiera de las siguientes cuatro bases nitrogenadas: Adenina (A), Guanina (G), Timina (T) y Citosina (C). El ADN es una doble hélice formada por dos polímeros lineales antiparalelos en forma de escalera, los "pasamanos" de la escalera son las moléculas de fosfato y desoxirribosa mientras que los "peldaños" son una de las cuatro bases anteriormente mencionadas, una de cada lado y unidas entre sí mediante puentes de hidrógeno.

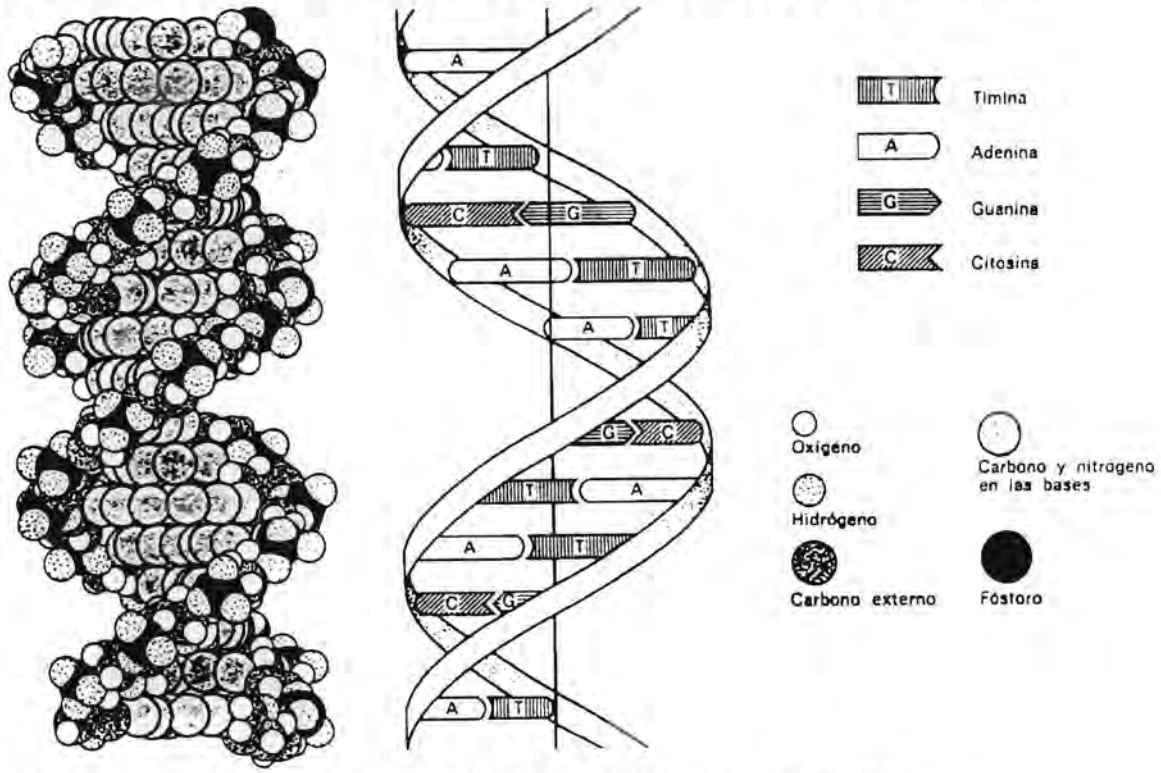


Figura 1. Esquema simplificado representando una doble hélice de ADN.

Adicionalmente, se debe mencionar que las bases nitrogenadas pueden ser de dos tipos: la A y la G pertenecen a la familia de las purinas y son moléculas de dos anillos ("grandes") (Figura 2); por otra parte, la T y la C pertenecen a la familia de las pirimidinas y son moléculas de un anillo ("pequeñas").

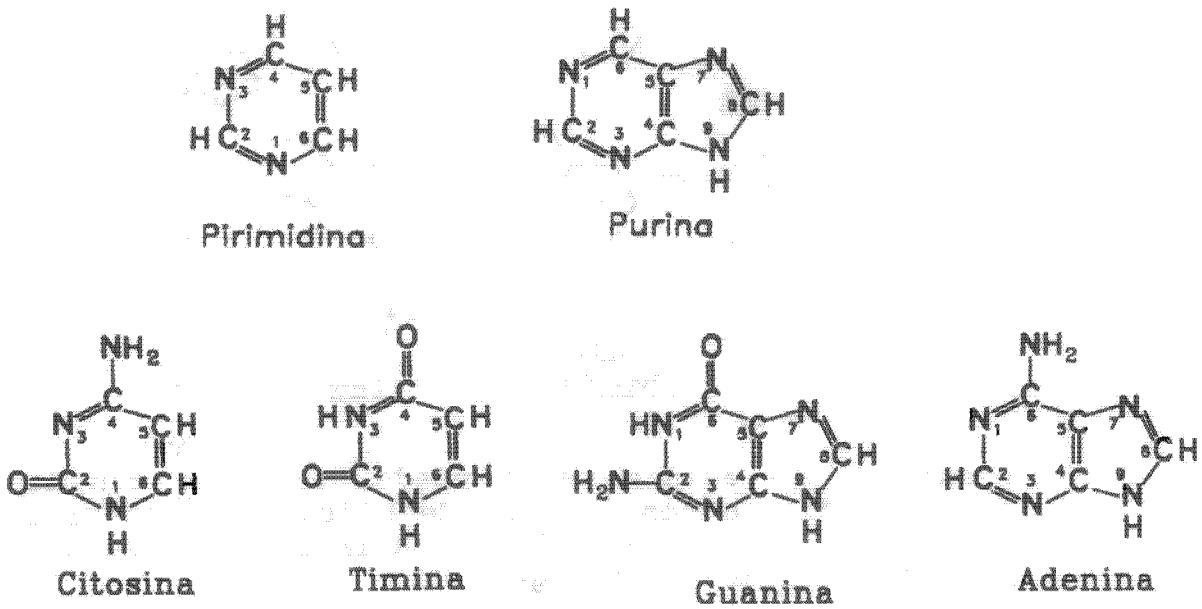


Figura 2 Las bases que componen la parte variable del ADN arregladas según su naturaleza purínica o pirimídica.

Como cada peldaño está formado por una base de cada lado, para mantener homogénea la anchura de la escalera es necesario que si en un lado de la doble hélice se encuentra una purina, del lado opuesto le corresponda una pirimidina. Esto, aunado al hecho (Figura 3) de que no todas las uniones purina-pirimidina son posibles debido a la incompatibilidad de formar adecuadamente los enlaces de hidrógeno, determina el siguiente principio de complementaridad entre los lados de la escalera: cada vez que de un lado aparece una molécula de adenina del otro lado existe

una de timina con la cual está aparejada y viceversa, mientras que cada vez que aparezca una de citosina se hallará aparejada con una de guanina y viceversa. Por ello, la información contenida en la molécula se puede deducir de una sola rama de la doble hélice.

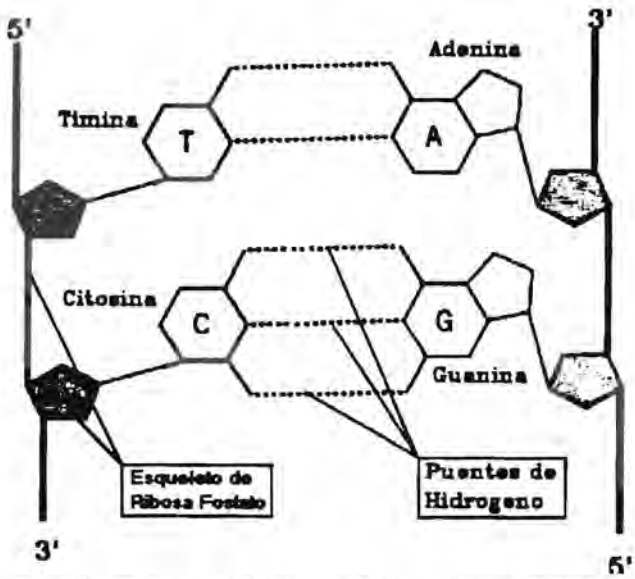


Figura 3. La adenina sólo puede aparearse con la timina mediante dos puentes de hidrógeno, mientras que la citosina sólo lo hace con la guanina mediante tres puentes de hidrógeno.

La unión de los peldaños de la escalera no es homogénea; las parejas A-T están unidas mediante dos puentes de hidrógeno, mientras que las parejas G-C lo están mediante tres puentes de hidrógeno (Figura 3). Esto sugiere que las secciones de la cadena más abundantes en G o en C sean, en promedio, más rígidas que aquellas en las cuales la abundancia relativa de A-T sea mayor, en el sentido de la energía necesaria para desnaturalizar (separar las dos ramas) la molécula. Esto, como se verá más adelante, no es necesariamente cierto. Por otra parte, la diferencia de tamaños entre las purinas y las pirimidinas hace que la escalera no sea geométricamente uniforme; si de un sólo lado se alternan purinas y pirimidinas la escalera se encontrará dislocada, de

manera que podemos afirmar que en este caso la molécula es, en promedio, rugosa. Si, localmente, se encuentran solamente purinas la cadena será lisa (en realidad la situación es mucho más complicada pero para nuestros propósitos esta descripción es suficiente). Para una descripción más amplia consultar [21]. De aquí en adelante, llamaremos a las bases G y C *fuertes* y a A y T, *débiles*.

La cadena de ADN contiene toda la información necesaria para la producción de cadenas de aminoácidos (polipéptidos) que, a su vez, dan lugar a las proteínas que son la materia prima tanto estructural como funcional de todo ser vivo. En el ADN se encuentra codificada la herencia morfológica[2] y funcional que una célula transmite a sus descendientes.

La doble hélice se abre como un cierre de cremallera, genéricamente hablando, para dos procesos diferentes de la vida celular. Uno de ellos es el proceso de replicación en el cual el ADN se duplica generando dos copias idénticas (salvo por mutaciones o errores de copiado) quedando una de estas copias en cada célula producto de la división celular. El otro proceso, llamado de transcripción-traducción (del cual volveremos a hablar adelante), es aquel mediante el cual la información contenida en el ADN determina la manera en que una célula produce los polipéptidos que a la larga darán lugar a las proteínas. Un segmento de ADN que contiene la información necesaria para la sintésis de una proteína se llama *gen estructural*.

Las células se clasifican como *procariontes* y *eucariontes*; las primeras no tienen un núcleo cuya pared separe al material

---

[2] Esto es una exageración. En realidad, una célula de embrión sufre un proceso de diferenciación que es respuesta no sólo del contenido genético (todas las células de un organismo poseen el mismo juego de genes) sino de mecanismos de contacto o interacciones de corto alcance entre las células ("información posicional", ver Wolpert [7]). Estos mecanismos podrían activar o desactivar genes reguladores que, como su nombre lo indica, regulan la actividad de los genes estructurales.

genético del resto del protoplasma, un ejemplo de ellas son las bacterias. Las células eucariontes, por otra parte, poseen un núcleo bien definido en el cual reside el ADN, todos los organismos pluricelulares están conformados por células eucariontes, en estas, no toda la molécula de ADN está constituida por genes; de hecho, estos conforman una mínima parte de la secuencia. Los genes están separados entre sí por largos segmentos intergénicos cuya función, si es que la tienen, no ha sido aún dilucidada. Aún más, un gen está constituido por uno o varios segmentos llamados *exones* que contienen la información que codifica la producción de polipéptidos y por islas entre ellos llamadas *intrones* cuya utilidad o falta de ella tampoco se conoce cabalmente. Al conjunto de material genético de una célula se le llama "genoma". El genoma humano consta aproximadamente de 2000 megabases (Mgb)[3] de las cuales, casi el cinco por ciento codifica para la producción de polipéptidos o regula el funcionamiento de esta producción; la función del resto, en caso de haberla, es desconocida.

La manera en que se lleva a cabo el proceso de transcripción-traducción es la siguiente: en algún momento de la vida celular el segmento de ADN correspondiente a un gen estructural (de aquí en adelante, simplemente *gen* mientras no se diga otra cosa) se abre como se mencionó anteriormente y en lugar de duplicarse en una molécula similar (Figura 4), la sucesión de bases se transcribe en una molécula de ácido ribonucléico (ARN)[4], conocido, en esta etapa, como ARN premensajero.

---

[3] Un virus tiene del orden de 10 Kb (Kb, mil bases), los animales y plantas más comunes tienen aproximadamente 2000 Mgb, algunos sapos y salamandras pueden llegar a tener 20,000 Mgb. Curiosamente, el organismo conocido que tiene el mayor número de bases en su genoma es un protozoario muy simple: la amiba *Chaos Chaos* con 1 Gb (un millón de Mgb).

[4] Siguiendo el principio de complementaridad mencionado arriba. De esta manera se obtiene una copia de una de las ramas de ADN, la única diferencia es que en el ARN se reemplaza la timina (T)

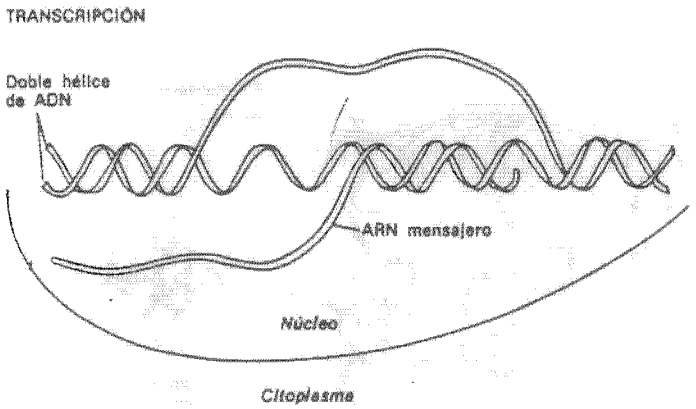


Figura 4. Esquema simplificado del proceso de transcripción de una secuencia de bases del ADN a una secuencia de ARN.

Después, en las células eucariontes, con ayuda de un sistema formado por proteínas y ARN, conocido como *espliceosoma*, se eliminan los segmentos intrónicos y se "pegan", en sucesión, las partes codificadoras; este proceso se conoce como *edición*. La molécula resultante, llamada ARN mensajero, viaja a una estructura de la célula conocida como *ribosoma*. En ésta, se lee secuencialmente el ARN (Figura 5) y por cada tres bases se agrega un aminoácido en una cadena lineal conocida como *polipéptido* y que es la precursora de una proteína. A cada terna de bases se le llama *codón*.

---

por uracilo (U), que también es una pirimidina (Figura 2) con propiedades geométricas y químicas muy semejantes a las de la timina.

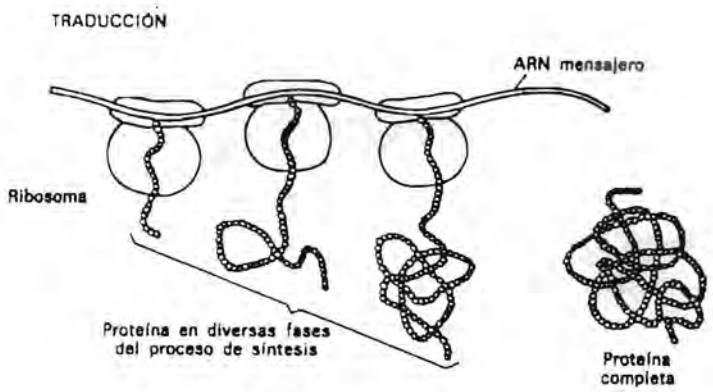


Figura 5. Esquema simplificado del proceso de traducción de una secuencia de bases de ARN a una secuencia polipeptídica.

Dado que existen cuatro bases, pueden formarse sesenta y cuatro codones distintos de los cuales todos, salvo tres, dan lugar a aminoácidos. Sin embargo, en la naturaleza existen tan sólo veinte aminoácidos diferentes que se emplean en la construcción de proteínas, de lo que se infiere que hay codones que, aunque distintos, son sinónimos para un aminoácido (Tabla 1). Esta redundancia o degeneración será determinante para definir las reglas de evolución de nuestro modelo.

1era posición	2da posición				3era posición
	T	C	A	G	
T	Fen	Ser	Tir	Cis	T
	Fen	Ser	Tir	Cis	C
	Leu	Ser	Alt	Alto	A
	Leu	Ser	Alto	Trp	G
C	Leu	Pro	His	Arg	T
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Tre	Asn	Ser	T
	Ile	Tre	Asn	Ser	C
	Ile	Tre	Lis	Arg	A
	Met	Tre	Lis	Arg	G
G	Val	Ala	Asp	Gli	T
	Val	Ala	Asp	Gli	C
	Val	Ala	Glu	Gli	A
	Val	Ala	Glu	Gli	G

Tabla 1. El Código genético. "Alto" es el codón de terminación del mensaje.

Durante los procesos anteriormente reseñados, puede suceder que, debido a un error en la duplicación o en la transcripción, se cambie una base por otra; a este fenómeno se le llama *sustitución*[5]. Existen varios tipos de sustituciones pero el más común es el que mantiene invariante el tamaño de la molécula mutada; es decir, que mantiene el carácter púrinico o pirimídico en la

---

[5] Hablaremos indistintamente, aunque no es completamente correcto, de "sustituciones", o bien, de "mutaciones".

sustitución. La fuente de las mutaciones puede ser de naturaleza muy diversa y no será tratada aquí. Por el momento basta señalar que una o varias alteraciones en la secuencia de bases pueden inducir alteraciones en los polipéptidos y, por ende, en las proteínas resultantes. Las mutaciones son una fuente de variabilidad en los organismos que les otorga la plasticidad necesaria para que éstos, en su momento, respondan ante presiones selectivas dando lugar a la evolución biológica.

Además de tener el ADN una estructura de doble hélice en el espacio, esta molécula tiene adicionalmente un enrollamiento sobre sí misma. Esto le confiere al ADN una gran complejidad espacial. Sin embargo, para los propósitos de este trabajo, interesan solamente las sustituciones que se dan en el proceso de duplicación o de transcripción, durante el cual el ADN (con ayuda de enzimas especializadas) se "estira" localmente, de manera que el suponer al ADN como una sucesión lineal de letras durante el mencionado proceso es una buena aproximación.

CAPITULO II

AUTOMATAS CELULARES

Los autómatas celulares (AC) fueron originalmente propuestos por Von Neumann y Ulam a finales de la década de los 40's. Su intención era mostrar que algunos fenómenos muy complejos tales como la reproducción, el crecimiento y la evolución de la vida misma podían ser estudiados como sistemas dinámicos y que sus características esenciales podían ser reducidas a un conjunto de reglas de interacción de elementos simples llamados *células*. La idea de Von Neumann, alentado por la necesidad de automatizar la producción fabril a causa de la escasa mano de obra durante la Segunda Guerra Mundial, era crear sistemas complejos capaces de reproducirse a sí mismos. Esta idea, más la sugerencia de Ulam de adoptar sistemas completamente discretos, dio origen a la Teoría de Autómatas Celulares. Esta teoría estudia estructuras homogéneas discretas, distribuidas espacialmente en arreglos llamados "células" ("celdas", "sitios"). Cada una de éstas puede adoptar un número finito de estados y el arreglo en su conjunto evoluciona temporalmente de manera discreta de acuerdo a una regla homogénea[6] en el espacio y de naturaleza local. Esto último se refiere a que la regla de evolución es una función cuyos argumentos son el estado particular de una célula y los estados de una colección de celdas vecinas al tiempo t y su imagen es el estado de la célula original al tiempo $t+1$.

---

[6] Esta condición no es estrictamente necesaria. De hecho, existen AC cuyas reglas pueden depender tanto del tiempo como de la posición. Sin embargo, en este trabajo se considerará como condición necesaria para definir un AC el hecho de que la regla sea homogénea; a tales AC se les puede llamar *Clásicos* si fuera necesario enfatizar que su regla es homogénea.

Los autómatas celulares más simples son aquellos en los cuales el arreglo espacial es unidimensional. Si este es el caso y etiquetamos a cada célula o celda con un número, entonces el valor de la variable que reside en la celda i al tiempo t se denota por $x_i(t)$ y sólo puede tomar valores en un conjunto discreto. Si la evolución del AC se realiza de tal manera que el estado de cada sitio se asigna sólo una vez que los valores de todos los sitios de la generación anterior se han asignado ya, entonces se dice que el AC es *Sincrónico*, de otra manera el AC es *Asincrónico*.

Los autómatas celulares unidimensionales se pueden clasificar, a su vez, atendiendo a la naturaleza del conjunto de sus valores. Bajo este criterio, los más simples son los AC *binarios* en los cuales, como su nombre lo indica, la variable sólo puede tomar los valores cero o uno. El hecho de que los AC evolucionen discretamente en el tiempo, siguiendo reglas locales y homogéneas se puede expresar como

$$x_i(t+1) = \mathbf{F}(x_{i-1}(t),\ x_i(t),\ x_{i+1}(t)) \qquad \forall\ i,t.$$

En este caso, se dice que la regla de evolución es de *primeros vecinos*, o bien, que es de *rango* 1. El AC puede ser *Determinístico* o *Probabilístico* dependiendo de la naturaleza de la función **F**[7]. Si se considera a los AC probabilísticos, aún se

---

[7] También se puede hablar de AC *reversibles* o *irreversibles*, dependiendo si **F** es invertible o no. Se puede demostrar [27] que un AC de segundo orden

$$x_{t+1} = \mathbf{F}(x_{(i),t}) - x_{t-1}$$

es un AC reversible para cualquier **F** ($\{i\}$ denota una vecindad). La importancia de estos AC radica en que pueden usarse como modelos de fenómenos de la Física Clásica. Debido al hecho de que en biología reproductiva se depende sólo de la generación anterior y al hecho de que en este trabajo se usarán reglas probabilísticas, aquí ya no se hablara de este tipo tan importante de AC.

puede hacer una distinción más fina: aquellos cuya regla de evolución sea probabilística propiamente hablando y aquellos cuya evolución es la aplicación con distintas probabilidades de dos o más reglas determinísticas [24].

Una regla de AC está plenamente especificada cuando se da el valor de $x_i(t+1)$ para todas las posibles configuraciones de entrada (vecindades). Para un AC binario y de primeros vecinos, el número de configuraciones de entrada es $2^3=8$ (las posibles permutaciones de dos objetos en tres lugares). Una regla se especifica si se dan todos sus valores posibles (entradas). Por ejemplo

$$\begin{array}{llll} 000 \rightarrow a_0 & 001 \rightarrow a_1 & 010 \rightarrow a_2 & 011 \rightarrow a_3 \\ 100 \rightarrow a_4 & 101 \rightarrow a_5 & 110 \rightarrow a_6 & 111 \rightarrow a_7 \end{array}$$

La regla está completa si se menciona la *Tabla de la regla*: $\{a_0, a_1, a_2, a_3, a_4, a_5, a_6, a_7\}$ o bien, un *Número de la regla* que es la representación decimal de la cadena binaria $(a_7, a_6, a_5, a_4, a_3, a_2, a_1, a_0)_2$. Una tabla o un número de regla contiene toda la información de un AC visto como una función. Sin embargo, a veces no se necesita toda esta información y basta con decir "más o menos" cuando dos reglas son parecidas o no. Para esto, se definen parámetros de *campo medio* [16]; con la regla de AC definida arriba, podríamos suponer que las entradas a_1, a_2 y a_4, de alguna manera son "semejantes" pues las tres están asociadas a vecindades que contienen dos ceros y un uno. Si se definen los parámetros n_1, n_2, n_3 y n_4:

$n_1 \equiv$ número de bits distintos de cero en $\{a_1, a_2$ y $a_4\}$
$n_2 \equiv$ número de bits distintos de cero en $\{a_3, a_5$ y $a_6\}$
$n_3 \equiv$ en $\{a_0\}$
$n_4 \equiv$ en $\{a_7\}$

En lugar de una tabla de regla, tenemos un conjunto $\{n_1, n_2, n_3, n_4\}$ de *campo medio* para la regla del AC. Más adelante se usará en este trabajo una aproximación de campo medio.

Los autómatas celulares son sistemas dinámicos completamente discretos y su relevancia reside en el hecho de que pueden usarse como modelos de procesos naturales en los cuales la naturaleza del fenómeno sea intrínsecamente discreta; esto proporciona una valiosa alternativa ante la costumbre tradicional de llevar el estudio de cualquier fenómeno al ámbito de lo continuo. Los autómatas celulares se han usado como herramienta matemática para modelar el crecimiento de cristales [9], la formación de patrones de coloración en animales [10], la formación de patrones geométricos en reacciones químicas [11], así como modelos de turbulencia química y modelos de hidrodinámica [12].

Aunque los autómatas celulares determinísticos poseen una gran riqueza dinámica, descrita por Wolfram [3], en este trabajo no serán utilizados. En efecto, la naturaleza estocástica de las mutaciones puntuales sugiere el uso de un autómata celular con regla de evolución probabilística.

En resumen, los cuatro aspectos que definen un autómata celular, son los siguientes:

i) Una estructura espacial discreta.
ii) Rango discreto de valores de las variables.
iii) Evolución temporal discreta.
iv) Reglas locales, homogéneas, de evolución.

Es importante tener en mente estos cuatro puntos pues una de las intenciones del presente trabajo es definir un AC a partir de consideraciones fenomenológicas. Más tarde se apelará a la memoria del lector en el momento en que se proponga un esquema que cumpla los cuatro puntos y se afirme que estamos en presencia de un AC.

Siendo los autómatas celulares funciones, en el sentido estricto del término, se ha intentado clasificarlos en familias. Entre estos intentos, se puede destacar el trabajo de Wolfram [3] que propone una clasificación derivada del comportamiento dinámico de los AC en estrecha analogía con la clasificación de los puntos de equilibrio de una ecuación diferencial. También se puede citar a Miramontes [24] que hace una clasificación "fenotípica" de los AC unidimensionales y a Gutowitz [26]. Consultar también el trabajo de Packard y Li [27].

CAPITULO III

EL MODELO

1.- DISCUSION.

Dado que la molécula de ADN es un polímero lineal, puede ser visto como un arreglo discreto en una dimensión. Cada sitio correspondería al lugar ocupado por una de las cuatro bases posibles. Así, este arreglo puede contener sólo cuatro posibles valores por celda o sitio (dos si usamos la representación degenerada F-D o 0-1). Si tomamos en cuenta que el ADN evoluciona discretamente en el tiempo sacando copias de sí mismo y eventualmente cambiando el valor de alguna celda mediante una mutación puntual o sustitución, hemos de reconocer que se satisfacen las primeras tres -de las cuatro- condiciones mencionadas al final del capítulo anterior. Debe ser claro en este momento que el único ingrediente que hace falta para poder identificar esta molécula con un AC es un conjunto de reglas locales de evolución que sean capaces de reflejar la enorme complejidad biológica y física de la evolución molecular. Esta similitud sugirió a Burks y Farmer [13] el uso de los autómatas celulares para modelar la evolución del ADN. Su trabajo es una descripción de las características estructurales y funcionales del ADN que lo hacen buen candidato a ser modelado por un AC[8]. Sin embargo, no se proponen reglas de evolución y sólo se menciona como importante la necesidad de escoger adecuadamente un criterio de selección para que se pueda hablar de *incorporación al genoma* de los eventuales resultados de una simulación. Hasta donde nosotros sabemos, no existe ningún reporte en la literatura de

---

[8] Justamente lo que se argumenta en las primeras líneas de este capítulo.

algún trabajo que haya propuesto algún modelo convincente de evolución molecular del ADN haciendo uso de los autómatas celulares[9]. El hecho de estar tratando con moléculas complejas que interactúan entre sí de manera también compleja y que contienen información, tanto en la sucesión de sus componentes como en su estructura misma, hace que esta tarea sea también un sistema complejo. Para desentrañar esta madeja y poder llegar a proponer un modelo completo, se comenzará por el estudio de las interacciones de los componentes del ADN.

2.- EL CODIGO GENETICO Y LA DEGENERACION DE LOS CODONES.

En la Tabla 1 se encuentra el código genético, esto es, la identificación entre tripletes de bases y aminoácidos. Si llamamos XYZ a las posiciones de las bases en un codón cualquiera, se conoce el hecho de que la posición Z es la menos importante para definir un aminoácido en el sentido de que es muy probable que si se altera la tercera letra de un codón, el aminoácido resultante sea el mismo. De igual manera la posición Y es determinante para la definición de un aminoácido ya que si se altera mediante una transversión (una sustitución que cambia el carácter púrinico-pirimídico de la base mutada) la base en esta posición, no solamente cambiará el aminoácido resultante, sino también la naturaleza polar de éste y esto influye fuertemente en la estructura y función protéica. La posición X queda en una situa-

---

[9] Existe un trabajo de Cocho y Martínez-Mekler [30], en el cual se propone un modelo con la misma intención que la del presente. Ellos usan un mecanismo de redes de mapeos acoplados. Para definir una Red de Mapeos Acoplados (CML) en su versión más simple, basta tomar lo cuatro puntos que definen a un AC clásico y cambiar la exigencia de valores de la variable en un conjunto discreto por la de variables de rango continuo. Por supuesto que esto lleva a situaciones extremadamente interesantes, por desgracia fuera del ámbito del presente trabajo.

ción intermedia. Todo esto se traduce en el hecho de que existan probabiblidades diferentes para que una mutación al azar se fije al genoma. En efecto, las mutaciones en la tercera posición -como seguramente no afecta la secuencia peptídica- tendrán una alta probabilidad de pasar a la progenie. Apliquemos razonamientos análogos a las demás posiciones del codón y si denotamos por P_z (P_x y P_y se definen análogamente) la probabilidad de que una mutación en la tercera posición sea incorporada al genoma, podemos resumir la discusión mediante la desigualdad:

$$P_z > P_x > P_y$$

La probabilidad diferencial de mutación dependiendo de la posición de una base en el codón, resulta de la restricción selectiva sobre la funcionalidad de los péptidos resultantes. A este hecho le llamamos *Selección Natural Externa* (SNE).

3.- ENERGIA DE ENLACE, RESTRICCIONES GLOBALES Y SELECCION NATURAL INTERNA.

Recientemente [14], se ha reportado en la literatura la energía de amarre entre parejas consecutivas de bases (*digramas*), dado que dos bases consecutivas y sus complementarias forman un paralelepípedo, la energía de la que se habla se debe entender como un promedio entre las seis interacciones resultantes. Estos resultados se encuentran en la Tabla 2.

DIGRAMA	ΔH(RNA)	ΔH(DNA)	TIPO
AA	6.6	9.1	DD-00
AT	5.7	8.6	DD-00
TA	8.1	6.0	DD-00
TT	6.6	9.1	DD-00
CC	12.2	11.0	FF-11
CG	8.0	11.9	FF-11
GC	14.2	11.1	FF-11
GG	12.2	11.0	FF-11
AC	10.2	6.5	DF-01
AG	7.6	7.8	DF-01
TC	13.3	5.6	DF-01
TG	10.5	5.8	DF-01
CA	10.5	5.8	FD-10
CT	7.6	7.8	FD-10
GA	13.3	5.6	FD-10
GT	10.2	6.5	FD-10

Tabla 2. Valores absolutos de la entalpía (ΔH) para digramas tanto de RNA como de DNA, los valores están dados en Kcal/mol. La cuarta columna indica el tipo de digrama en la representación fuerte-débil (0-1).

Es claro que, a partir de esta Tabla, se puede afirmar que la energía de enlace no depende de las bases particulares, sino que depende solamente de su carácter fuerte o débil : '1' y '0', respectivamente. De manera que se puede afirmar que, aproximadamente

para ARN:

$$\Delta H(11) > \Delta H(10,01) > \Delta H(00) \qquad (1)$$

y para ADN:

$$\Delta H(11) > \Delta H(00) > \Delta H(10,01) \qquad (2)$$

Esta pareja de desigualdades nos dice que tanto en el ADN

como en el ARN la energía de amarre de parejas de bases consecutivas no se distribuye homogéneamente sino que depende del *tipo de digrama* que puede ser fuerte-fuerte, débil-débil, fuerte-débil y débil-fuerte.

Por otra parte, se conoce el hecho[10] [15] de que en el ARN correspondiente a exones de células eucariontes sucede que

$$\Delta H_{zx} > \Delta H_{yz} > \Delta H_{xy}$$

Esta desigualdad nos dice, a su vez, que la energía de amarre, en exones, tampoco se distribuye homogéneamente pero ahora dependiendo de la *posición del digrama dentro del codón.* Dicho de otra manera, la tercera posición del codón se une a la primera del codón siguiente con una energía de amarre mayor que la segunda posición con la tercera y ésta, a su vez, lo hace con una energía mayor que la de la unión de la primera base del codón con la segunda. De aquí se desprende la existencia de una periodicidad (periodo 3) en la distribución de las energías de amarre. En efecto, la energía va como ...D-R-F-D-R-F-... en donde D, R y F significan, débil, regular y fuerte, respectivamente.

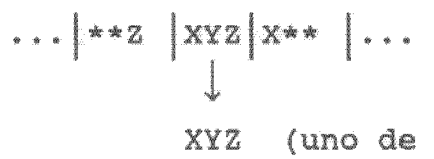
Estas inhomogeneidades en la distribución local de energía imponen restricciones fisicoquímicas en la estructura de los ácidos nucléicos pues si ambas se han de cumplir simultáneamente, una secuencia de bases no puede tener una abundancia homogénea de los ocho tipos posibles de codones (en la representación 0-1). A estas restricciones les llamaremos *Selección Natural Interna* (SNI).

---

[10] Esto se hizo estadísticamente, considerando directamente cada digrama y buscando la energía asociada en la Tabla 1 y relacionándola con su posición en el codón. Todo esto se hizo para varios genes del banco GenBank.

4.- LA REGLA DE EVOLUCION DEL AUTOMATA.

Los aspectos fenomenológicos mencionados arriba permiten hacer una propuesta realista de regla para el autómata celular. La regla tiene que respetar tanto la SNE -representada por las diferentes probabilidades de fijación de una mutación al genoma- como las desigualdades (1) y (2) que se derivan de la SNI. La propuesta que se hará rompe un poco la idea natural de que el mecanismo de autómata celular actúe a nivel de bases, en efecto, aquí se propone que tal mecanismo opere a *nivel de codones*. Es decir, el elemento básico que defina una variable discreta será el codón. Dado que en nuestra representación sólo existen ocho codones (000, 001, 010,...,111), el conjunto en el cual el AC toma valores es este espacio de codones. Con estas ideas, un autómata de primeros vecinos puede tener 8^3 posibles vecindades por sitio (dos celdas "tías" y una celda "madre"). Sin embargo, la función local que nos dirá -en el párrafo siguiente- a que codón puede mutar un codón dado como función de su vecindad tendrá que respetar las desigualdades (1) y (3) de la SNI y, por lo tanto, sólo interesa la posición Z del codón vecino por la izquierda y la posición X del codón vecino por la derecha. Gráficamente se tiene el esquema de transición siguiente:



Aquí el símbolo * representa cualquier base. Bajo estas condiciones la regla del autómata consta solamente de 32 entradas en lugar de las 8^3 predichas. Para decidir qué codón particular es el resultado de una vecindad específica, se propone

una función que evalúe qué tanto (o qué tan poco) respeta una mutación puntual la desigualdad que se deriva de la inhomogeneidad en la distribución de la energía dependiendo de la posición del digrama en el codón[11]. Esta función tiene un valor inicial de 3; penaliza con un punto a cada violación que un posible codón resultante haga de esta desigualdad y premia con un punto a un codón por cada vez que se respete la desigualdad, como sólo existen cuatro digramas en una vecindad del AC, esta función tiene como conjunto imagen a {0,2,4,6}. A manera de ejemplo podemos ilustrar las posibles transiciones del codón 000 teniendo como vecino izquierdo al codón **1 y como vecino derecho al codón 0**

**1	000	0**
4	100	
0	010	
6	001	

Para construir esta tabla, se tomó en cuenta que el valor promedio de la entalpía para las uniones 0-0 es de 6.75, para las uniones 1-0 ó 0-1 es de 10.4 y, por último, para las uniones 1-1 es de 11.65. Así, este ejemplo muestra que una mutación en la primera posición recibe una calificación de 4 puesto que la vecindad 1|100|0 consta de las uniones 1-1 (con amarre 11.65), 1-0 (con 10.4) y dos 0-0 (6.75). De esta manera la función diría (usando las desigualdades (3)): "...11.65 es mayor que 10.4; un

---

[11] Puesto que la Selección Natural Externa es aquella que se deriva de las restricciones sobre el péptido resultante, para incluirla en la regla del AC se toman las desigualdades (1) y no las (2), dado que estas restricciones actuarían a nivel de ribosoma y, por tanto, sobre el ARN. La regla resultante del segundo caso sería distinta y no viene al caso. También, con los mismos propósitos, para las consideraciones que involucren cálculos de las entalpías de amarre, se toman las correspondientes a la columna de ARN.

punto a favor y llevamos 4 de calificación; 10.4 no es menor que 6.75 por lo que quito un punto y llevamos 3; 6.75 es menor o igual a 6.75 por lo que agrego un punto y llevamos 4 y ésta es la calificación final...". En particular la regla nos señala que el codón 000 con esta vecindad, da lugar al codón 001. En este momento, es pertinente recordar que la SNE proporciona la probabilidad de que una mutación se fije al genoma, de manera que la entrada de la regla que corresponde a esta situación, se ve de la siguiente manera

$$**1|000|0** \rightarrow 001 \text{ con probabilidad } P_z.$$

La tabla íntegra de transiciones se muestra en el Cuadro 2. Los empates en las calificaciones se resuelven generando un número al azar en el intervalo [0,1] y eligiendo la mutación de acuerdo con la probabilidad que le toque por su posición en el codón.

Aunque la propuesta de regla de evolución del autómata está completa, para darnos una idea de su dinámica, se hace una aproximación de campo medio (Ver el capítulo II. Se hace un promedio de las imagenes de todas las entradas de la regla que provienen de la misma vecindad, en este caso del mismo codón). En la Figura 6 se muestra una representación gráfica de esta aproximación.

Se observa que el codón 001 es un *atractor fuerte*[12], mientras que el 011 y el 101 lo son *débiles*, los demás vértices del cubo son repulsores. Es interesante subrayar que el codón 001 codifica para ocho aminoácidos distintos y que entre los tres codones que son atractores codifican un total de 19 aminoácidos, es decir, que salvo por dos, estos tres bastan para re-

---

[12] En el sentido de que la probabilidad de que una mutación en los codones que son sus preimágenes es grande. Un atractor débil se define consecuentemente.

CUADRO 2

0

0 | 000 | 0		0 | 000 | 1		1 | 000 | 0		1 | 000 | 1	
100	2	100	2	100	4	100	4
010	0	010	2	010	0	010	2
001	4	001	4	001	6	001	4

1

0 | 001 | 0		0 | 001 | 1		1 | 001 | 0		1 | 001 | 1	
101	2	101	2	101	4	101	4
011	0	011	4	011	2	011	4
000	2	000	2	000	4	000	4

2

0 | 010 | 0		0 | 010 | 1		1 | 010 | 0		1 | 010 | 1	
110	4	110	2	110	0	110	2
000	2	000	2	000	4	000	4
011	0	011	4	011	2	011	4

3

0 | 011 | 0		0 | 011 | 1		1 | 011 | 0		1 | 011 | 1	
111	0	111	2	111	0	111	2
001	4	001	4	001	6	001	6
010	0	010	2	010	0	010	2

4

0 | 100 | 0		0 | 100 | 1		1 | 100 | 0		1 | 100 | 1	
000	2	000	2	000	4	000	4
110	4	110	2	110	0	110	2
101	2	101	2	101	4	101	4

5

0 | 101 | 0		0 | 101 | 1		1 | 101 | 0		1 | 101 | 1	
001	4	001	4	001	6	001	6
111	0	111	2	111	0	111	2
100	2	100	2	100	4	100	4

6

```
0| 110 |0        0| 110 |1        1| 110 |0        1| 110 |1
   010   0          010   2          010   0          010   2
   100   2          100   2          100   4          100   4
   111   0          111   2          111   0          111   2
```

7

```
0| 111 |0        0| 111 |1        1| 111 |0        1| 111 |1
   011   0          011   4          011   2          011   4
   101   2          101   2          101   4          101   4
   110   4          110   2          110   0          110   2
```

presentar íntegramente el código genético. Si se deja evolucionar solo a un AC con esta regla, cabría esperar que tienda a secuencias con una gran abundancia de 001, una abundancia menor de 101 y 011 y los demás codones representados en pequeñas cantidades.

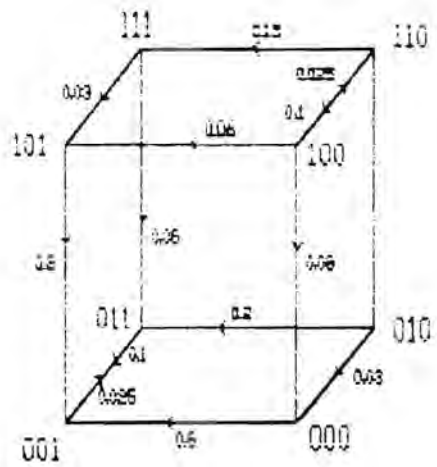


Figura 6. Cubo cuyos vértices son los ocho posibles codones y cuyos aristas representan posibles mutaciones. Los números indicados en cada arista representan las probabilidades de transición entre dos tipos de codones (vértices).

En este momento, se tienen ya los cuatro ingredientes para definir la molécula de ADN como un autómata celular. Sin embargo, queda pendiente la delicada tarea de decidir cuando un conjunto de mutaciones aumenta o reduce la adecuación de una secuencia de ADN y, por lo tanto, la hace candidata a fijarse al genoma o a desaparecer. Tal tarea se emprende en el siguiente capítulo.

5.- EL ESQUEMA FORMAL DEL MODELO.

1. Comenzar con una población de genes (secuencias aleatorias de longitud *N* en la representación fuerte-débil (alfabeto {0,1}).
2. Mutar cada individuo de acuerdo al esquema de AC.
3. Permitir que se reproduzcan aquellos individuos con alta adecuación y desechar a aquellos que no tengan suficiente adecuación (se define en el capítulo siguiente).
4. Regresar a 2.

CAPITULO IV

"Incorporating an appropriate notion of selection is one of the important unsolved problems in this field".
(Burks y Farmer, [13]).

1. EL FILTRO SELECTIVO

El criterio de adecuación para decidir cual cadena debe sobrevivir y cual no, debe tener fuertes bases fenomenológicas; con base en esta consideración debemos encontrar un índice o medida que indique qué tan cercana o lejana es una cadena de nuestra simulación a una real. Se ha elegido el índice D_{WS} propuesto por Cocho *et al* [17]. Tal índice se define como sigue

$$D_{WS} = \frac{N_{00}N_{11}-N_{10}N_{01}}{N_{10}N_{00}+N_{01}N_{11}}$$

donde N_{ij} es el número de digramas i,j con i y j en $\{0,1\}$.

D_{WS} mide el grado de agregación de digramas 0 y 1. Se puede comprobar que secuencias de ceros y unos distribuidos al azar y con composición 50-50 dan valores de D_{WS} cercanos al cero, que secuencias con los ceros y unos completamente alternados dan valores de D_{WS} cercanos al 1 y, por último, que secuencias con

grandes islas de ceros y grandes islas de unos dan valores de D_{WS} cercanos al -1. Como se discute en [17], los valores de D_{WS} no se distribuyen al azar en secuencias reales; todo lo contrario, las células eucariontes tienen valores de D_{WS} negativos y las procariontes cercanos al cero (ver Figuras 7 y 8). En [17] se define el índice D_{YR} de manera análoga a D_{WS} a partir de la representación binaria purina-pirimidina. Aunque aparece también en las Figuras 7 y 8, no se empleará en este trabajo.

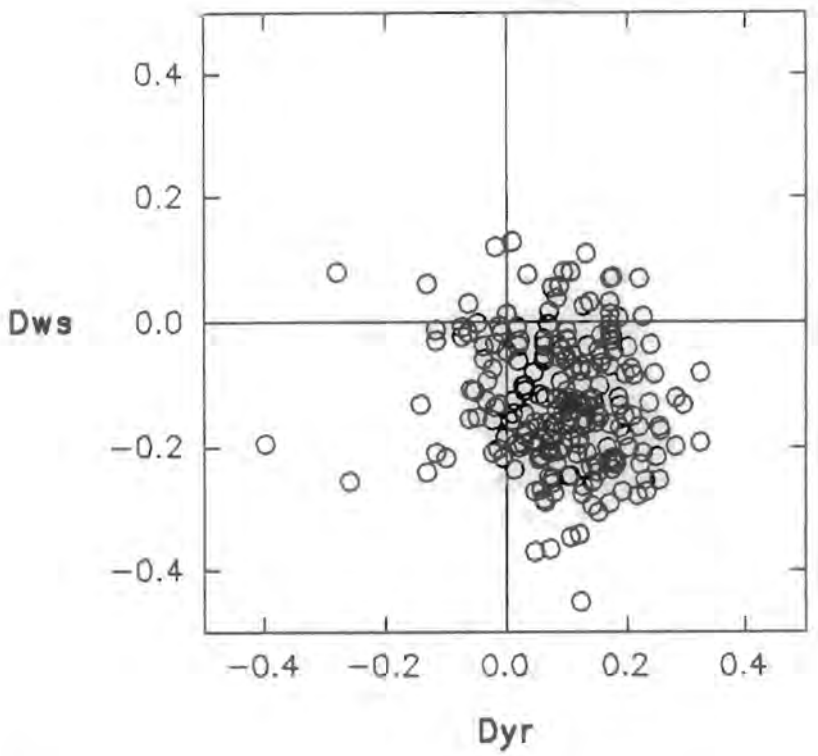


Figura 7. Diagrama Dyr-Dws para 247 secuencias eucariotes obtenidas en el GenBank.

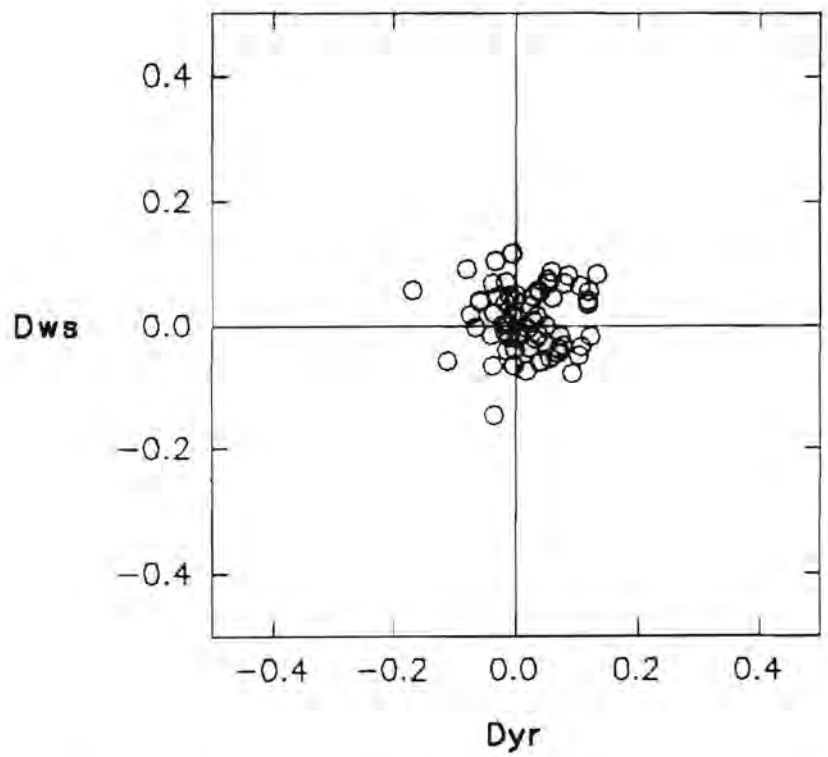


Figura 8. Diagrama D_{YR}-D_{WS} para 63 secuencias procariotes obtenidas en el GenBank.

Adicionalmente, la Figura 9 muestra el histograma[13] que se obtiene de dividir en intervalos el rango de D_{WS} para secuencias de eucariontes. Este histograma juega un papel fundamental en este trabajo. En efecto, nos servirá como base de comparación entre las secuencias que se obtengan de la simulación y las reales. En pocas palabras, de aquí se desprenderá el criterio de selección.

---

[13] 250 secuencias tomadas del GenBank. El rango se dividió en 15 clases, se normalizaron los valores al mayor y los valores del histograma se graficaron con línea continua con interpolación de esplín.

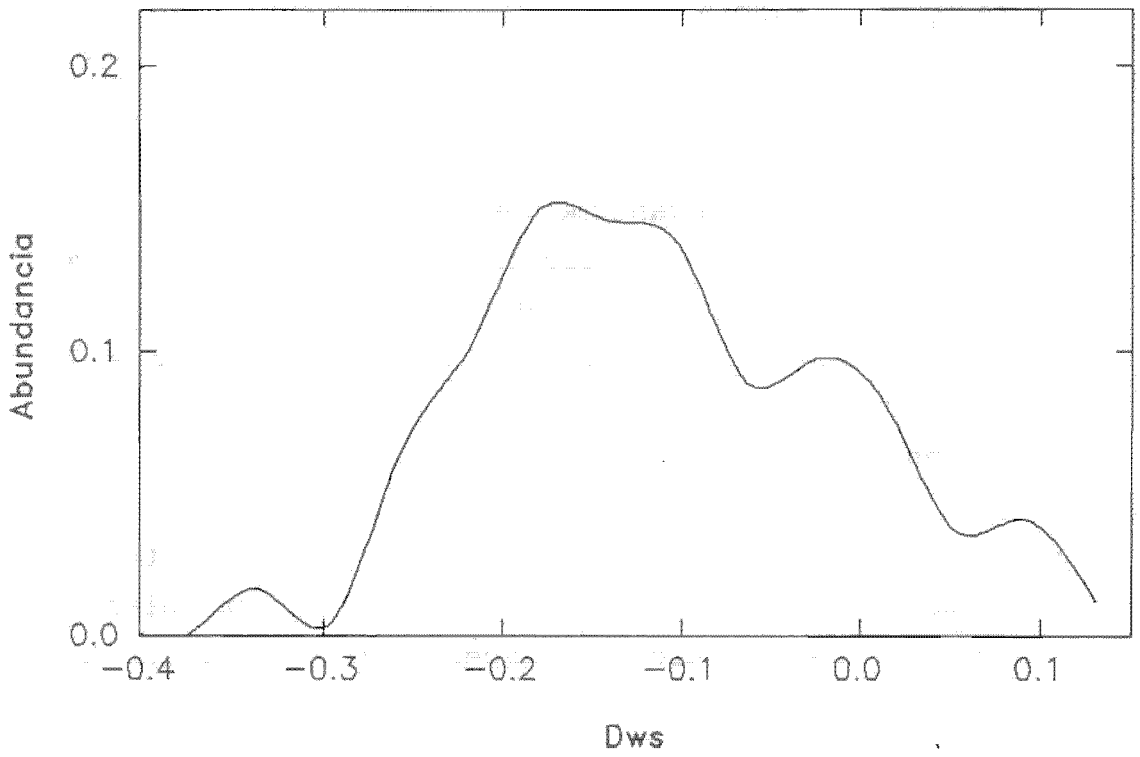


Fig 9. Histograma de la distribución de los valores de Dws de los puntos del eje vertical de la Figura 7 .

Tomaremos el índice D_{WS} como criterio de qué tan "eucarionte" es una secuencia. La dinámica hallada en el capítulo anterior tendería, por sí sola, a llevar cualquier secuencia inicial a una de valor D_{WS}=-0.5[14] y esto es un reflejo de la actuación individual del mecanismo de AC. Si, como se discute en

---

[14] La secuencia ...001001... no contiene digramas 11, por lo que N_{11} es cero. Por otra parte, en esta secuencia se puede ver que los demás digramas aparecen en las mismas proporciones; cada uno, una tercera parte. Si se efectúan los cálculos, Dws resulta valer menos un medio.

[17], se ha de tomar en cuenta la selección natural actuando sobre la secuencia de aminoácidos resultante después del paso por el ribosoma, la SNE debería forzar a una secuencia a tender hacia valores de D_{WS} cercanos al cero. De esta manera, el filtro selectivo que se propone aquí es muy sencillo de enunciar: toda mutación que haga descender el valor absoluto de D_{WS} incrementa la adecuación de una secuencia. De esta manera, nuestro modelo en su totalidad comprende un mecanismo de mutación local (interacción de corto alcance), asociado tanto a la selección natural externa como a la interna, en competencia con un mecanismo de selección global (restricciones de largo alcance) asociado a la selección natural interna y que tiende a disminuir el valor de D_{WS}. Como se mencionó en la introducción, la actuación de dinámicas en conflicto no es desconocida para las personas que estudian sistemas complejos; de hecho, se reconoce como una fuente de riqueza dinámica que evita los estados asintóticos planos. Considérese, a manera de ejemplo, el mapeo logístico discreto; una dinámica de crecimiento debida a la reproducción de los individuos (interacción de corto alcance) contrapuesta a una dinámica de castigo a la sobrepoblación (efectos globales, restricciones de largo alcance) dan como resultado (para valores adecuados del parámetro) una conducta compleja, caótica.

Resumiendo, el filtro selectivo que se propone aquí se puede esquematizar de la siguiente manera:

1. Medir D_{WS} en cada individuo (secuencia binaria).
2. Elegir aquellos que tengan los valores más pequeños de D_{WS} (en valor absoluto).
3. Permitir a los sobrevivientes reproducirse.

En este esquema, se reconoce el espíritu del método de optimización de los *Algoritmos Genéticos* [18,19,20].

2. LOS ALGORITMOS GENETICOS.

Siempre ha existido la necesidad de optimizar algo. Estamos mal acostumbrados a que ese "algo" venga dado en forma de una función y, así, el problema de optimizar es el de encontrar máximos o mínimos de esa función. Los métodos más socorridos son los derivados del Análisis Numérico. Sin embargo, esta clase de métodos imponen serias restricciones a la clase de funciones que se han de optimizar, por ejemplo, las funciones han de ser continuas, lisas, los extremos han de ser unimodales, etc. Las funciones que no tienen estas características se salen del ámbito del Análisis Numérico pero están lejos de carecer de interés teórico y práctico [20]. Una alternativa a este problema la constituyen los métodos heurísticos, de entre estos, sobresalen los *Algoritmos Genéticos* (AGs) [20]. En los AGs se mantiene como paradigma de optimización a la evolución biológica; se parte del hecho de que el mecanismo de la selección natural elige a los individuos más aptos sobre los menos aptos, a los primeros les permite reproducirse y copiar a su descendencia las características que los hacen aptos y de esta manera la población incrementa globalmente la medida de su aptitud. Como método de búsqueda de extremos de una función, los AGs han probado su calidad al ser empleados en multitud de problemas que escapan a las posiblidades del Análisis Numérico y al usarse, con buenos resultados, en los problemas que pueden ser resueltos por el Análisis Numérico. El esquema general de los AGs puede resumirse de manera algorítmica [18]:

El algoritmo opera sobre un conjunto $B(t)$ de M cadenas $\{c_1, c_2,..,c_M\}$ sobre el alfabeto $\{0,1\}$ con adecuaciones (valor a maximizar) $A_j=A(c_j,t)$. La secuencia de pasos a seguir es la siguiente

1.- t=0.

2.- Se calcula la adecuación promedio $A(t)$ de todas las cadenas en $B(t)$ y se normaliza la adecuación de la cadena j a $A(c_j,t)/A(t)$.

3.- Se asigna a cada cadena c_j en $B(t)$ una probabilidad proporcional a su valor normalizado. Usando esta distribución de probabilidad, se eligen M cadenas de la muestra inicial (algunas, las de mayor adecuación tienen más oportunidad de aparecer repetidas, las de menor adecuación tienen pocas posibilidades de aparecer en la nueva muestra).

4.- A las cadenas seleccionadas, se les aplica un *operador genético*. En otras palabras; una fuente de variabilidad, que puede ser un mecanismo de mutaciones puntuales, traslocaciones (intercambio de segmentos entre dos secuencias apareadas al azar; un mecanismo primitivo de sexualidad [8]), inversiones (cambiar el orden de un segmento dentro de una secuencia, de manera que el principio se vuelva fin y al revés), etc. En este modelo, la fuente de variabilidad está dada por el mecanismo de autómata celular propuesto anteriormente. No obstante, continuaremos hablando de *mutaciones*.

5.- Ponga t=t+1.

6.- Regrese a 2.

La descripción anterior amerita algunos comentarios. En el punto número 3 se propone *una manera* de elegir a los individuos más aptos para que tengan mayor probabilidad de dejar descendencia. Esta manera no es, ni con mucho, la única. Cualquier forma de privilegiar reproductivamente a los más aptos es válida; imagine el lector, a manera de ejemplo, que seleccionamos a los, digamos, 10 "mejores" individuos, que eliminamos al resto y que restituimos la población a su tamaño original mediante copias de

los elegidos. Esto sería una selección "elitista", pues los individuos de baja adecuación ni siquiera tienen la esperanza de que la ruleta probabilística los elija en su ciego girar. El lector podrá hacer propuestas de selecciones "democráticas", "sectarias", etc.

CAPITULO V

RESULTADOS Y CONCLUSIONES

1. LA SIMULACION.

En este trabajo se eligió como proceso de optimización el reseñado en la última sección del capítulo anterior. Los algoritmos genéticos constituyen una alternativa interesante para algún trabajo futuro en el cual se quiera incorporar al modelo dinámica cromosómica más complicada: traslocaciones y entrecruzamiento, principalmente. Esto es una propuesta para un modelo de evolución de secuencias intrónicas.

Para la simulación se escogieron cadenas iniciales binarias, azarosas, en la representación 0-1, con una longitud de 150 bases (parecida en longitud a las regiones exónicas reales en células eucariontes). Se permitió una población máxima de 200 secuencias. Las probabilidades de mutación se tomaron en la proporción 2:1:6 para los sitios X,Y y Z, respectivamente. Estos datos provienen de las proporciones de mutación reales de los genes de la β-globina humana [21].

Como ya se mencionó, el mecanismo de AC actuando por sí solo, lleva las secuencias al punto $D_{WS}=-0.5$, para implementar la dinámica antagónica, se forzó a que el algoritmo genético maximizara la función $f(D_{WS})=a(D_{WS}+1)^{k}$[15], ver la Figura 10, en donde los parámetros a y k se ajustaron mediante una búsqueda de Montecarlo hasta lograr que el máximo de la población en el his-

---

[15] Esta función es monótona creciente y tiene su único cero en $D_{WS}=-1$. Al no estar acotada, la búsqueda se su máximo hace que el algoritmo genético empuje siempre hacia la derecha, mientras que el autómata celular lo hace hacia la izquierda.

tograma D_{WS} coincidiera con el obtenido de las secuencias reales. La forma de esta familia de funciones, permite elegir mediante el parámetro k, la *fuerza* con la que actúa la dinámica que se opone al autómata celular. El parámetro a hace ajustes finos.

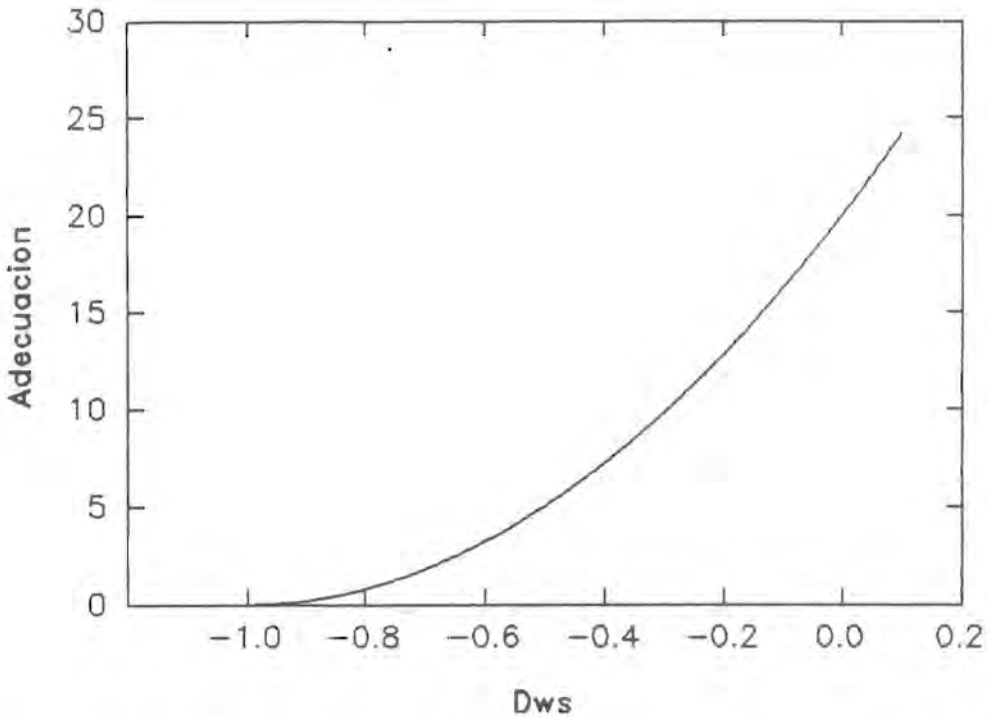


Figura 10. Forma de la función de adecuación propuesta $f(D_{WS})=a(D_{WS}+1)^k$.

Para poder estimar el número necesario de iteraciones, se permitió al sistema correr por 250 generaciones y se calculó en cada paso el promedio del índice D_{WS} para todos los individuos de la población. Los resultados de esta corrida exploratoria se muestran el la Figura 11. Se puede apreciar que con 100 iteraciones se está bastante cerca de un comportamiento parecido al asintótico. No se puede -ni se debe- afirmar que este sistema eventualmente alcanzará un equilibrio asintótico, pues por la

construcción misma del modelo lo más probable es que las soluciones muestren el fenómeno de frustración[16].

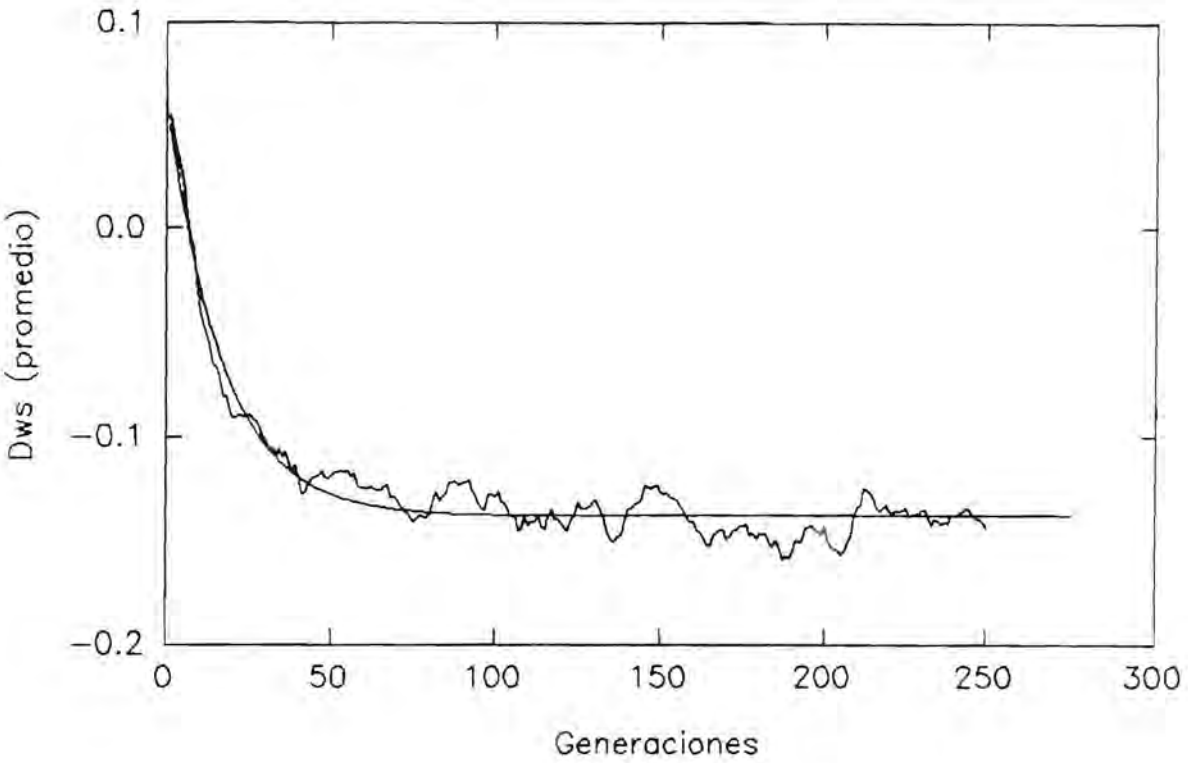


Figura 11. Promedio de los valores de Dws para todos los individuos de la población como función del tiempo (generaciones). La linea continua es la regresión a: $a + b \exp(c\ D_{WS})$.

---

[16] A menudo se observa en sistemas discretos, con pocos estados posibles por elemento, el fenómeno de *Frustración.* Imaginemos un triángulo equilátero que tiene en cada vértice un elemento que solo puede adoptar dos estados posibles ("más" o "menos", por decir algo). Supongamos que la ley de interacción local nos dice que cada elemento tendrá que estar en un estado diferente al de sus vecinos próximos. Si el elemento del primer vértice está en "mas", el del segundo tendrá que estar en "menos" y cualquier estado del tercero violará al menos una de las interaciones. Un sistema cuyas interacciones no pueden ser satisfechas, todas a la vez, de manera simultánea, se llama *Frustrado.*

2. LOS RESULTADOS.

En la Figura 12 se muestra el histograma D_{WS} de la población inicial, como es de esperarse es una distribución aproximadamente simétrica, centrada en el origen de coordenadas. En la Figura 13, se muestra la misma población después de 100 generaciones.

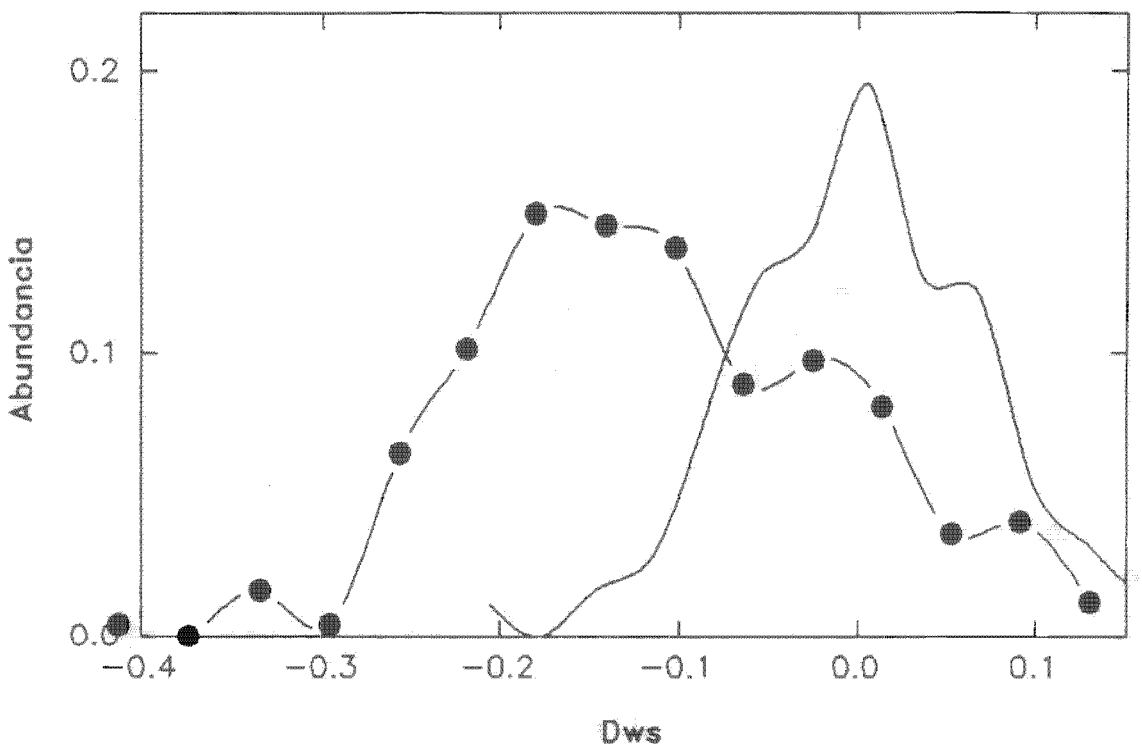


Figura 12. Histograma de los valores de Dws de una población cuyos individuos fueron generados al azar. La línea cortada representa al histograma de la población real.

La simulación se realizó con los valores a=20 y k=2 (determinados por búsqueda Montecarlo). Se puede apreciar que estos valores en los parámetros hacen coincidir los máximos de ambas distribuciones y no el rango de dispersión. Esto puede deberse a algún mecanismo natural de aleatoriedad no contemplado en el presente modelo. De hecho, se sospecha que una restricción adicional no contemplada en el presente trabajo, sea la asociada al índice D_{YR}. Actualmente, se piensa en un modelo que contemple esta posibilidad[17]. El área bajo las curvas no es la misma pues el número de secuencias es diferente para cada caso.

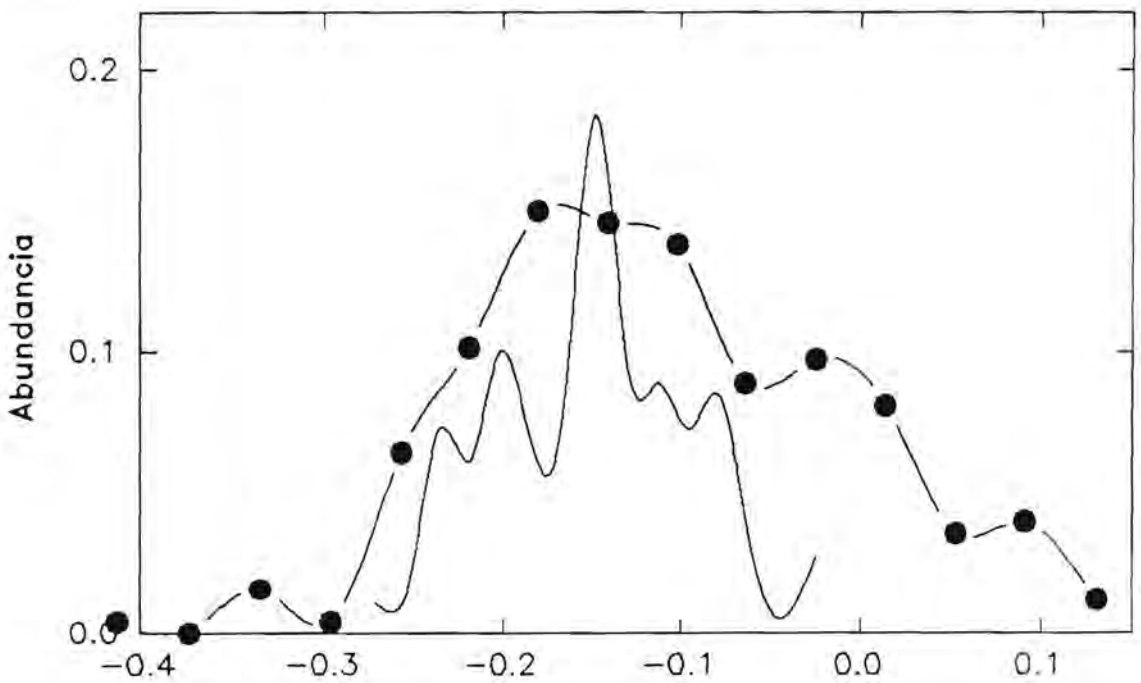


Figura 13. Histograma de los valores de Dws de la población de la Figura 12 después de 100 generaciones. La línea cortada representa al histograma de la población real.

---

[17] Este futuro modelo es más complicado que el presente, pues para poder incluir dicho índice se tiene que abandonar la representación 0-1 y trabajar en la representación de las cuatro letras, lo que complica enormemente la deducción de la regla del autómata celular.

En la Figura 14 se muestra el histograma que enseña la abundancia relativa de cada uno de los ocho codones en secuencias reales. La Figura 15 contiene el histograma de abundancia de codones para la población de secuencias resultante de la simulación. Es claro que nuestro modelo reproduce cualitativamente este aspecto de la naturaleza de las secuencias exónicas pues en todos los casos, salvo en el codón siete, las proporciones relativas entre los codones de secuencias reales y las simuladas, coinciden.

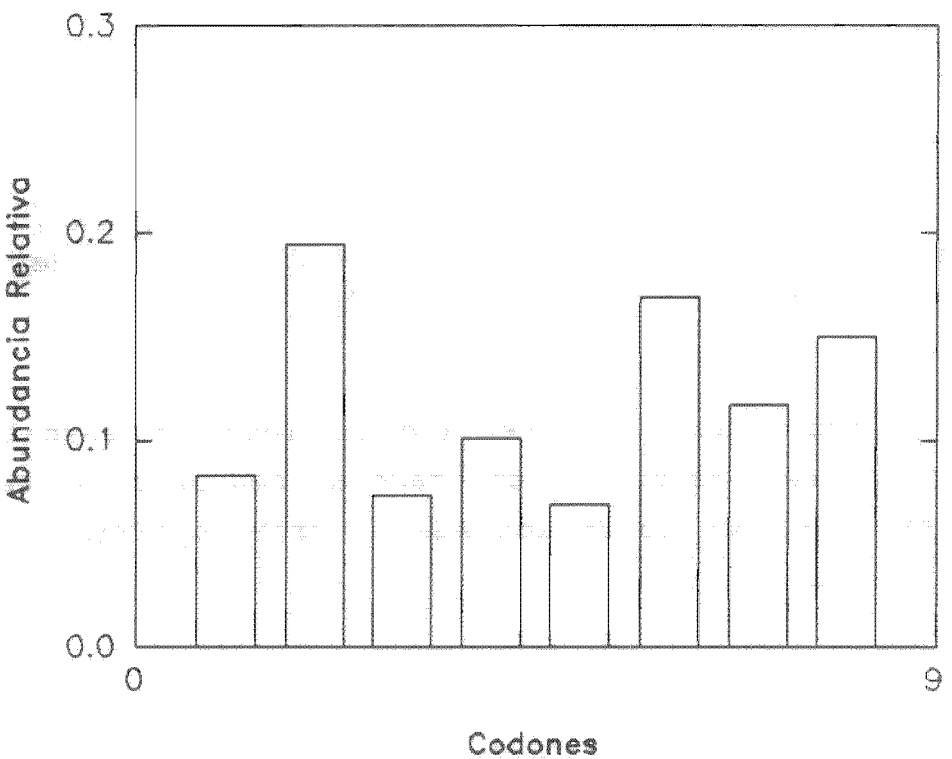


Figura 14. Abundancia relativa de los ocho tipos de codones en genes reales. Las columnas corresponden a los ocho tipos diferentes de codones, ordenados del cero al siete.

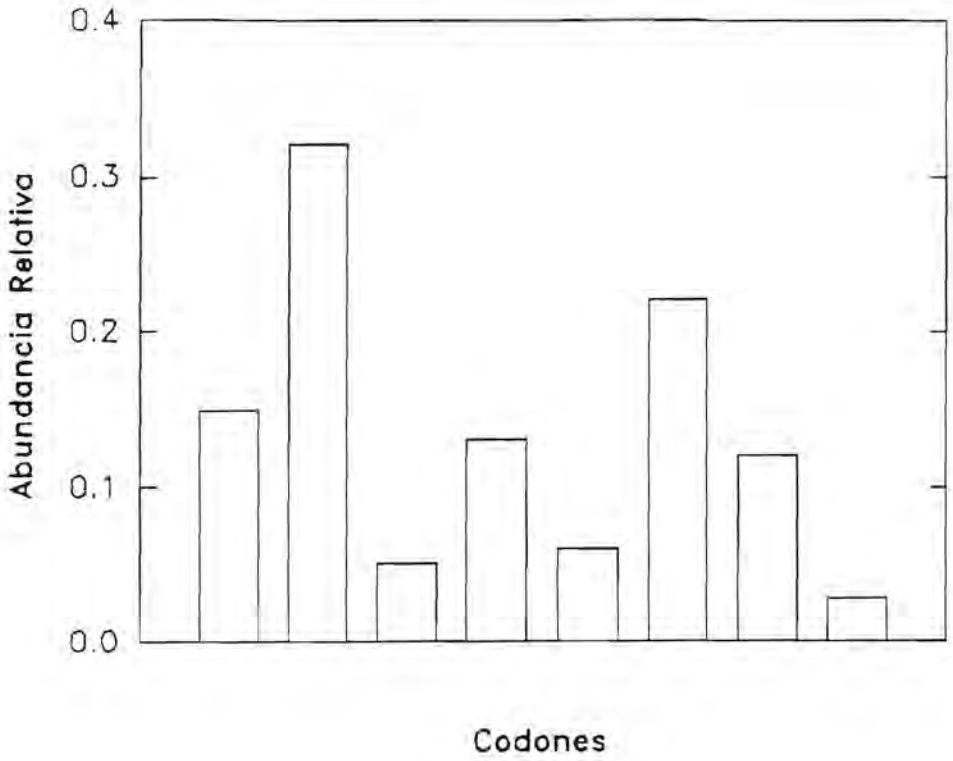


Figura 15. Abundancia relativa de los ocho tipos de codones en la población simulada. Las columnas corresponden a los ocho tipos diferentes de codones, numeradas del cero al siete.

Toda la programación se realizó en lenguaje Pascal. Las simulaciones se realizaron en computadoras con procesadores 286 y 386 de Intel. El código de los AGs se probó optimizando funciones con extremos conocidos.

3. CONCLUSIONES.

Un primer resultado de este trabajo, además de la simulación y sus resultados, lo constituye la filosofía global que subyace en el modelo; esto es, la trascendencia que puede tener el hecho de concebir la estructura y función del ADN desde el punto de vista de su naturaleza fisicoquímica y sujeta a restricciones de tipo termodinámico.

El modelo reproduce características cuantitativas medibles en las secuencias reales, lo que sugiere que comparte características esenciales con el mecanismo real de la evolución biológica de las secuencias genéticas. Esto trae consigo consecuencias interesantes:

1. Algunos investigadores [30] han propuesto la hipótesis del "ADN basura" (junk DNA). Sucintamente, se pueden resumir tales ideas de la siguiente manera: no obstante que sólo una pequeña parte del ADN (ver capítulo I) codifica para aminoácidos, todo el ADN en conjunto, se replica para la reproducción de las células. Esto condujo a algunas personas a pensar que los mecanismos de adaptación selectiva actuarían a nivel de la replicación del genoma, en otras palabras, que la unidad mínima de selección evolutiva, no sería ni la especie, ni la comunidad, ni el individuo sino, justa y precisamente, el genoma. A tal idea se le denomina *Selección por gen egoista*. En consecuencia, el ADN que no codifica es "basura". De acuerdo a nuestro trabajo, la selección opera sobre el genoma completo, pues aunque haya sectores que no codifican, forman parte de la estructura sobre la cuál la selección opera a manera de restricciones estructurales (Capítulo II).
2.- Desde el punto de vista de la dinámica, este trabajo utiliza la noción de *Sistemas Complejos* como método de trabajo y como marco ideológico[18]. En efecto, los resultados del trabajo muestran la riqueza conceptual y metodológica de términos tales como *soluciones de frustración, dinámicas en conflicto,* etc.

---

[18] Pudiera ser que a algunas personas les moleste el término. Si ese es el caso del lector, úsese como sinónimo de *Paradigma* o de *Conjunto de Ideas Dominantes.*

3.- De este trabajo se desprende la posibilidad de enriquecer la teoría de los algoritmos genéticos mediante la introducción de nuevos operadores genéticos tales como los autómatas celulares. Hasta donde sabemos, el uso de los AC en estas circunstancias es original y valdría la pena continuar trabajando en esta dirección.

4.- Otro aporte del presente trabajo es la sugerencia, ya mencionada, de incorporar las restricciones estructurales que se derivan de la alternanacia purina-pirimidina para hacer más restrictivo el filtro selectivo.

5.- Por último, se puede también realizar este trabajo en la representación de las cuatro bases, validando los resultados mediante algún algoritmo "reconocedor" de secuencias exónicas como el de Fickett [23]. Asimismo, sería interesante introducir algún índice de "complejidad informativa" para estudiar si ésta se incrementa a lo largo de la simulación (sería de esperarse), podría emplearse la medida propuesta por Lipmann [29].

BIBLIOGRAFIA

[1]. R. May. Simple mathematical models with very complicated dynamics. *Nature*, **261** (1976) 459-467.

[2]. P.W. Anderson. More is different. *Science* **177** (1972) 393.

[3]. S. Wolfram. Universality and complexity in cellular automata. *Physica* **10D** (1984) 1.

[4]. S. Wolfram. Statistical mechanics of cellular automata. *Rev. Mod. Phys.* **55** (1983) 601.

[5]. G. Cocho and J.L. Rius. Structural constraints and gene dynamics. *Rivista di Biologia-Biology Forum*. **82** (1989) 344.

[6]. D. Stein. Spin Glasses. *Sci. Am.* **261** (1989) 36-42.

[7]. L. Wolpert. *Curr. Top. Dev. Biol.* **6** (1971) 183-224.

[8]. Origins of Sex. L. Margulis and D. Sagan. Yale University Press. New Haven (1986).

[9]. N. Packard. Lattice models for solidification and aggregation. Proceedings of the first international symposyum for science on form. (Tsukuba, Japan. 1985).

[10]. G. Cocho, R. Pérez-Pascual, J.L. Rius and F. Soto. Discrete systems, cell-cell interactions and color pattern of animals. *J. Theor. Biol.* **125** (1987) 419.

[11]. B. Madore and W. Freedman. Computer simulation of the Belousov-Zhabotinsky reaction. *Science*, **222** (1983) 615.

[12]. Theory and applications of cellular automata. S. Wolfram (ed). (World Scientific, 1986).

[13]. C. Burks and D. Farmer, *Physica* **10D** (1984) 157-167.

[14]. K.J. Breslauer, R. Frank, H. Blöcker and L.A. Marky, *Proc. Natl. Acad. Sci. USA*, **83** (1986) 3746-3750.

[15]. G. Cocho and J.L. Rius. En: Theoretical Biology, eds. B.

Goodwin and P. Saunders. (Edinburgh University Press, 1989).
[16]. W. Li. Phenomenology of non-local cellular automata. Santa Fe Series 91-01-001.
[17]. G. Cocho, L. Medrano, P. Miramontes and J.L. Rius. Selective Constraints over DNA Sequence. En "Biologically inspired Physics". Plenum Press. 1991.
[18]. L.B. Booker, D.E. Goldberg and J.H. Holland. *Artificial Intelligence* **40** (1989) 235.
[19]. J.H. Holland. Genetic algorithms and the optimal allocation of trials. *SIAM J. Comput.* **2** (1973) 88.
[20] Genetic Algorithms in search, optimization and machine learning. D.E. Goldberg. Addison Wesley. Reading, Mass. (1989).
[21] J. Efstratiadis, J. Posakony, T Maniatis, R. Lawn, C. O'Connell, R. Spritz, J. DeRiel, B. Forget, Sh. Weissman, J. Slightom, A. Blechl, O. Smithies, F. Baralle, C. Shoulders y N. Proudfoot. The structure and evolution of the human β-globin gene family. *Cell* **21** (1980) 653-658
[22] L. Medrano and P. Miramontes. Constraints over DNA sequence. *Rivista di Biologia-Biology Forum.* **84** (1991) 9-11.
[23] J. Fickett. *Nucl. Acids. Res.* **10** (1982) 5303-5317.
[24]. Algunos aspectos de la teoría de autómatas celulares y sus aplicaciones en Biofísica. O. Miramontes. Tesis de licenciatura en Física, Facultad de Ciencias, UNAM. 1988.
[25]. N. Margulus. Physics-like models of computation. *Physica* **10D** (1984) 81.
[26]. H. Gutowitz. Hierarchical classification of cellular automata. *Physica* **45D** (1990) 136-156.
[27]. N. Packard y W. Li. Structure of the elementary cellular automata rule space. *Complex Systems* **4** (1990) 181-207.
[28]. G. Cocho y G. Martínez-Mekler. On a coupled map lattice formulation of the evolution of genetic sequences. *Physica* **51D** (1991) 119.

[29] D. Lipmann y W. Wilbur. Contextual constraints on synonymous codon choice. *J. Mol. Biol.* **163** (1983) 363-376.
[30]. L. Orgel y F. Crick. Selfish DNA: the ultimate parasite. *Nature* **284** (1980) 604-607.