Un Modelo Biológico en la Informática: Los Algoritmos Genéticos

José María Barreiro Sorribas
CETTICO-Fundación General UPM
Facultad de Informática Universidad Politécnica de Madrid
Campus de Montegancedo s/n 28660 Boadilla de Monte - Madrid
email: jmbarreiro@fi.upm.es

Felix Montañés Pazos
IBM. Madrid.
email: fmontanez@vnet.ibm.com

Resumen

El estudio de la biología puede ayudar al informático, y viceversa, a orientarse y entender algunos de los términos empleados con frecuencia en este dominio. No sólo es interesante como un dominio de investigación para la Informática, sino que también proporciona un conjunto rico de metáforas para ponderar la inteligencia: algoritmos genéticos, redes de neuronas y autómatas darwinianos, son una muestra reducida de los enfoques computacionales al comportamiento basado en los conceptos biológicos. Los informáticos y los biólogos tienen en muchos casos expectativas muy distintas acerca de la colaboración, la educación, los congresos y muchos otros aspectos, aparentemente banales de la investigación. Para trabajar con biólogos, los investigadores informáticos (en especial los de Inteligencia Artificial), deben poseer un conocimiento bastante profundo del dominio y buscar modos de acercar estas culturas científicas diferentes.

En este trabajo se verán, en primer lugar, los conceptos biológicos que han servido como base al desarrollo de modelos computacionales (citología, genética, bioquímica, biología molecular y evolución), se hará

un repaso de los modelos que constituyen lo que se ha dado en llamar Computación Evolutiva, viendo algunos de los modelos biológicos que han sido aplicados en la informática y, finalmente, las contribuciones de la informática en la biología haciendo especial hincapié en la genética.

1.- Conceptos Generales

Tanto en la biología como en la matemática y la informática, muchas de las percepciones verdaderamente brillantes provienen del reconocimiento de que dos puntos de vista aparentemente distintos son congruentes. Durante los últimos 50 años, los biólogos se han preocupado cada vez más por la base molecular de la función biológica en sus intentos de responder a preguntas del tipo de cómo saben las células cuando tienen que dividirse, cómo funciona la memoria o cómo ven los mamíferos.

1.1.- Citología

Cada uno de los 75 billones de células del hombre es una estructura viva que puede sobrevivir indefinidamente, y la mayor parte de casos incluso reproducirse siempre que los líquidos del medio contengan nutrientes adecuados. Las diferentes sustancias que componen la célula se denominan en conjunto protoplasma. Éste incluye, principalmente, las cinco substancias básicas siguientes: agua, electrólitos (potasio, magnesio, fosfato, sulfato, y carbonato y pequeñas cantidades de sodio y cloruro de calcio), proteínas (estructurales y enzimáticas), lípidos (triglicéridos, fosfolípidos y colesterol) y carbohidratos (glucosa y glucógeno o polímero insoluble de la glucosa).

En todos los organismos excepto en las bacterias, las algas azules y los virus, los cromosomas de cada célula están ordinariamente confinados dentro de una región limitada por una membrana denominada núcleo. Estos organismos se denominan nucleados o eucariotas, en contraste con las bacterias y algas azules que son procariotas. Las células eucarióticas presentan un núcleo en el interior del cual está alojado el material genético o cromosomas. Cuando el núcleo no está en fase de división o mitosis se denomina núcleo de interfase y en él no es posible identificar a los cromosomas como entidades individuales sino que aparecen como una red amorfa llamada cromatina [Creighton, 1983].

1.2.- Genética

La información genética poseída por un individuo es lo que constituye su genoma. Los portadores físicos de la información genética tanto en los organismos nucleados como en los anucleados son los cromosomas, cada uno de los cuales es portador de un subconjunto distinto de genes y, por tanto, de distintas secuencias de nucleótidos.

Los genes controlan la herencia de padres a hijos y la función diaria de la célula y su reproducción. Aunque el número de genes existentes en un núcleo de célula humana es desconocido, se calcula que los genes controlan la formación de, por lo menos, varios miles de tipos diferentes de proteínas esenciales para las funciones de las diversas células [Guyton, 1984]. Es, por tanto, esencial para la supervivencia de las especies que cuando una célula da lugar a dos células hijas, cada célula reciba una colección completa de cromosomas. Esta distribución equitativa cromosómica corre a cargo de un proceso de división celular llamado *mitosis.*

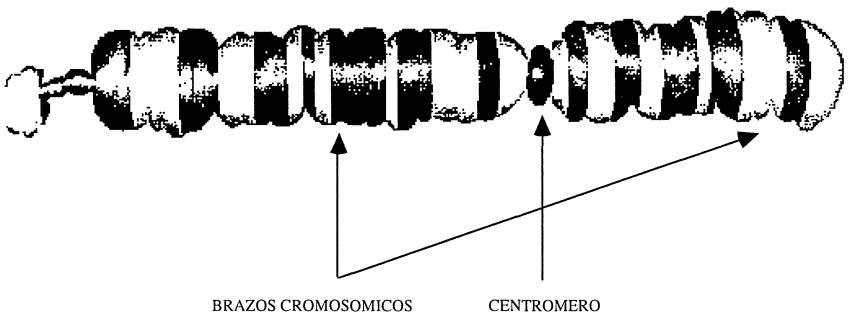

Figura 1-. Cromosoma acrocéntrico

En la célula humana en metafase aparecen un total de 46 pares de cromosomas hermanos. Estos difieren unos de otros en cuanto a tamaño y en la posición del centrómero. Así, se clasifican en metacéntricos cuando el centrómero está situado en posición media y por tanto los brazos del cromosoma son exactamente iguales; acrocéntricos (figura 1) cuando un

brazo es más largo que el otro; el resto de tipos de cromosomas son intermedios entre estos dos tipos.

Una descripción completa y ordenada, según su tamaño, de todos los pares de cromosomas que posee un tipo celular dado, constituye su cariotipo (figura 2). En el caso del ser humano existen dos representantes de cada tipo por lo tanto presenta 23 tipos diferentes de pares de cromátidas. Los cromosomas numerados del 1 al 22 se llaman autosomas para diferenciarlos de los cromosomas X e Y o cromosomas sexuales (X para la mujer e Y para el hombre).

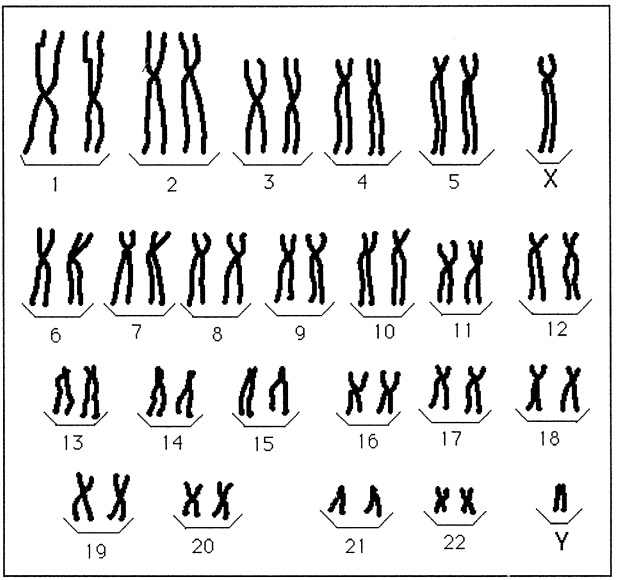

Figura 2.- Cariotipo.

En un cariotipo cada tipo cromosómico aparece dos veces, es lo que se denomina un número diploide, a diferencia del número haploide en los cuales sólo está presente un representante de cada tipo de cromosoma, como ocurre en el caso del óvulo y del espermatozoide. La dotación cromosómica haploide contiene toda la información del genoma y cabe

esperar, por tanto, que contenga un mínimo de una copia de cada tipo de genes poseídos por el progenitor [Goodenough, 1981].

1.3.- Bioquímica

Los bioquímicos descubrieron que una clase muy amplia de moléculas, llamadas proteínas, son responsables de casi todas las funciones del organismo vivo. Además, se observó que todas las proteínas se componían de una cadena lineal de elementos llamados aminoácidos. De hecho, los bioquímicos determinaron que existen sólo alrededor de 20 tipos de aminoácidos de la vida. Típicamente, se conectan entre sí de 50 a 500 de estos aminoácidos para formar una proteína. De manera más significativa, los bioquímicos observaron que estas proteínas, una vez constituidas, se doblaban para constituir formas únicas a partir de su secuencia de aminoácidos [Lehninger. 1972].

1.4.- Biología Molecular

La Biología molecular ha facilitado un enlace entre la genética y la bioquímica. La biología molecular es la disciplina que demostró la relación entre los genes y las proteínas cuando se determinó en 1953 [Watson, 1953] que el gen se compone de DNA (figura 3).

Figura 3.- Estructura tridimensional de una cadena de DNA.

Los genes actúan determinando los tipos de proteínas elaborados por las células. El DNA no es el molde directo para esta síntesis pues éstos están constituidos por moléculas de RNA (ácido ribonucleico). El flujo de la información genética en la mayoría de las células es, según se muestra en la figura 4:

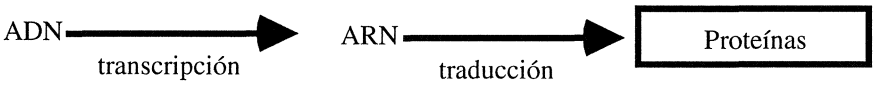

Figura 4.- Dogma central de la biología.

Ciertas moléculas de RNA son las portadoras intermediarias de la información en la síntesis de proteínas, mientras que otras moléculas de RNA son parte de la maquinaria de dicha síntesis. Con la aparición de las técnicas de ADN recombinante se puede clonar, secuenciar, manipular y expresar el ADN lo que ha dado lugar a una nueva industria: La Biotecnología. ADN, ARN y proteína, formas codificadas de la información genética, se adaptan bien a su manipulación por computadora.

1.5.- Evolución

Es el resultado de dos procesos: la selección natural (los que sobreviven hasta reproducirse) y la reproducción (mezcla y recombinación de genes en la descendencia). Cuando se habla de evolución hay que tener en cuenta dos conceptos principales:

- Recombinación: Dos cromosomas intercambian trozos de material genético
- Mutación: Alteración de una parte del material genético

2.- Computación Basada en Modelos Biológicos

Los modelos que se desarrollan en la investigación científica tienen varias formas. Algunos se construyen como objetos concretos (como una maqueta de un coche o el modelo de un circuito hidrodinámico de un sistema vascular), otros son más abstractos y existen sobre el papel o incluso como un programa de ordenador. Un modelo es una representación de una cosa (sea verbal, simbólica, matemática o física) como otra. En este contexto, encontramos que los modelos en un dominio del conocimiento se emplean como una fuente para los modelos en otro dominio diferente.

Los dominios o disciplinas de investigación (computación y biología) comparten áreas comunes que podrían realizar una aportación, y animar e inspirar a los investigadores que proceden de una amplia gama de especialidades científicas para que se interesen por este área y que comuniquen sus ideas. La forma y la función biológica ha sido durante mucho tiempo una fuente de inspiración y motivación para la gente que la ha empujado a construir nuevos artefactos, en muchos casos muy interesantes. Esto no sólo incluye aproximaciones como:

- Redes de neuronas artificiales (RNA)
- Algoritmos inmunes y sistemas de clasificación
- Algoritmos evolutivos (que incluyen algoritmos genéticos y estrategias de evolución)

Sino también ideas motivadoras de conceptos como:

- Autómatas celulares
- Matrices sistólicas
- Vida artificial
- Hormigas virtuales
- Ecologías computacionales.

El gran interés que existe en este campo de investigación es porque los sistemas biológicos proporcionan una gran variedad de conceptos y de ejemplos que abordan muchos de los temas y problemas de la informática. Así, entre otros, destacan:

- Procesamiento paralelo distribuido
- Ecologías de sistema abierto
- Comportamiento adaptativo no lineal
- Sistemas autónomos (autoorganizativos)
- Interacción, cooperación y comunicación
- Procesamiento asíncrono
- Evolución del software

- Emergencia y organización

La computación basada en modelos biológicos considera como fuente de sus modelos las ideas de las ciencias biológicas e intenta transferir las ideas relevantes a temas computacionales. En este proceso, existe un traslado de los conceptos de un dominio (biología) a otro (informática). Se infiere, de lo mencionado previamente, que existen dos campos distintos, pero estrechamente relacionados, para la investigación:

[A] Computación basada en modelos biológicos, es decir, temas computacionales que utilizan fuentes biológicas.

[B] Biología basada en modelos informáticos, es decir, temas biológicos que utilizan fuentes informáticas.

Por tanto, está pues claro que un diálogo interactivo entre los profesionales de las ciencias de la vida y los informáticos podría proporcionar un valioso medio de clarificar las ideas y los conceptos y también sugerir nuevas direcciones multidisciplinarias para la colaboración investigativa.

La metáfora describe el cerebro o la mente como un tipo de computadora y las ideas asociadas con el proceso de la información han sido trasladadas del campo de la informática y de la cibernética al dominio de las ciencias de la vida. Éste es un ejemplo de la biología/psicología basada en modelos informáticos. Sin embargo, de alguna forma, la metáfora de "red de neuronas" ha pretendido describir un elemento de software/hardware como un tipo de estructura cortical cerebral, es decir, las ideas procedentes de la neurofisiología han motivado desarrollos en la informática.

SISTEMAS BIOLÓGICOS COMO FUENTES INFORMÁTICAS:	
Sistemas a nivel subcelular, como bombas de iones y la organización de las proteínas en las membranas.	Ideas sobre la organización emergente y los factores implicados en este proceso. (organización en términos de entropía e información).
Modelo neo-darwiniano: extensiones y alternativas que den cabida a ideas como: - la fluidez genómica dentro de un ciclo de vida - la transcripción inversa - la amplificación y conversión de genes - los rasgos lamarkianos	RNAs, los algoritmos genéticos, los algoritmos inmunes, las estrategias de evolución Reto: desarrollar modelos en los que el código pueda responder autoadaptándose al entorno (no a través de mutación aleatoria y la recombinación)
Sistemas de modulación nerviosa y endocrina relacionados con la comunicación y la integración.	- Las RNAs son análogos pobres - El algoritmo de retropropagación es un ejemplo básico - La implicación multisistema de las interacciones químicas entre el sistema nervioso y el sistema inmunológico
Biotecnología.	Sistemas de inmovilización, máquinas biomoleculares y bioreactores
Las glándulas (incluyendo el cerebro y el hígado)	Ordenadores químicos (de datos) que procesan de forma altamente paralela

Sistemas simbióticos, la interacción y la comunicación unicelular e intercelular (célula única y mixta)	Estudios teóricos y empíricos de la naturaleza y de la interacción de agentes autónomos

2.1.- Computación Evolutiva

El increíble universo de la computación evolutiva constituye una pequeña muestra de lo que supone el universo científico que incorpora Sistemas Difusos, Redes de Neuronas Artificiales y que son denominadas "computación inteligente", lo que a su vez supone una pequeña parte de lo que incorporan cada vez más avances científicos: Vida Artificial, Geometría Fractal y otras ciencias de sistemas complejos y que constituyen la Computación Natural. Es decir, según el grado de complejidad de la computación biológica tenemos (figura 5):

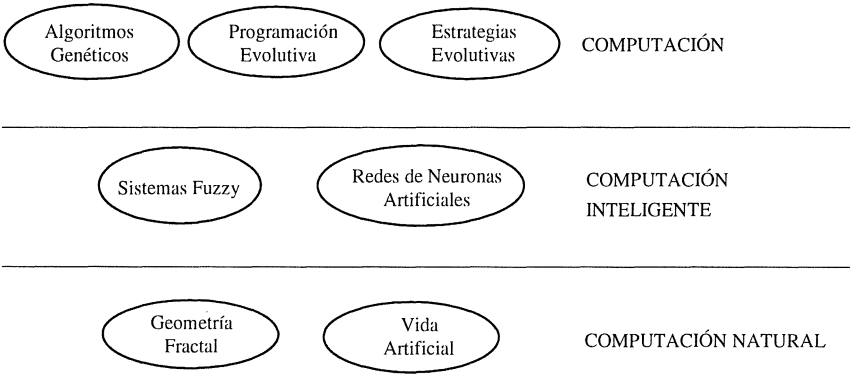

Figura 5.- Niveles de complejidad de la computación biológica.

En los últimos años los algoritmos de optimización global imitan ciertos principios de la naturaleza proporcionados en varios dominios de aplicación, representado principalmente por aquellos principios en los que la naturaleza ha encontrado "islas estables" en un "océano turbulento" de posibles soluciones. Los procesos biológicos han dado lugar a los siguientes métodos de optimización: Redes de Neuronas Artificiales y el campo de la Computación Evolutiva (CE).

En la CE se distinguen las siguientes ramas:

- Programación Evolutiva (PE)
- Estrategias Evolutivas (EE)
- Algoritmos Genéticos (AG)
- Sistemas Clasificadores (SC)
- Programación Genética (PG)

Otras muchas estrategias para solucionar problemas están basados sobre observaciones biológicas inspiradas en el descubrimiento de Charles Darwin en el siglo XIX: el significado de la selección natural y la supervivencia de los más aptos o adaptados de su "Teoría de la Evolución". Los algoritmos que se inspiran en ésta han sido denominados Algoritmos Evolutivos (AE).

Los AE forman un amplio abanico de términos que son usados para describir los sistemas solucionadores de problemas basados en computadoras y que usan modelos computacionales de los procesos evolutivos como elementos clave en su diseño e implementación. Todos ellos muestran un concepto común basado en la simulación de la evolución de los individuos a través de procesos como la selección, mutación y reproducción. Este proceso va a depender de la adaptación de estas estructuras o individuos en lo que ha sido definido como "entorno".

Concretamente, los AE mantienen una población de individuos que evolucionan según las reglas de la selección y otros operadores llamados "operadores de búsqueda u operadores genéticos" como son la recombinación y la mutación. Cada individuo en la población reacciona a medida que se produce su adaptación al entorno. La reproducción pone atención a la alta adaptación individual explotando la información de la capacidad de adaptación. La recombinación y la mutación perturban a esos individuos proporcionando heurísticas generales para la exploración. Desde un simple punto de vista biológico estos algoritmos son suficientemente complejos para proporcionar un robusto y poderoso mecanismo de búsqueda adaptativa.

2.1.1.- Programación Evolutiva

La Programación Evolutiva es una estrategia de optimización estocástica parecida a los algoritmos genéticos, pero que prescinde de las representaciones "genómicas", así como la recombinación en sentido de estrategia de mutación.

Al igual que los algoritmos genéticos, la Programación Evolutiva es útil a fin de optimizar las soluciones de los problemas cuando se excluyen otras técnicas como la descendencia de gradiente o el descubrimiento analítico directo. La optimización de la función, en la que la superficie de optimización es desigual y posee muchas soluciones localmente óptimas, se adapta bien a la técnica de la Programación Evolutiva.

En el proceso para la Programación Evolutiva, al igual que ocurre con los Algoritmos Genéticos, existe un supuesto subyacente de que la "aptitud" puede caracterizarse en términos de variables y que existe una solución óptima en términos de esas variables.

El método básico de la Programación Evolutiva se compone de 3 pasos, que se repiten, hasta que se supere el umbral de iteración o se obtenga una solución adecuada:

1) Elección de una población inicial de casos de prueba de forma aleatoria. El número de casos tiene mucho que ver con la velocidad de la optimización, pero no se dispone de respuestas acerca de cuál es el número idóneo de casos.

2) Cada población se replica en una nueva población. Cada uno de estos individuos descendientes se mutan según una distribución de tipos de mutación, que va de menor a mayor en un espectro continuo de tipos de mutación.

3) Cada descendencia se evalúa por medio del cálculo de su aptitud. Se escogen los N individuos mejores o bien se escogen aleatoriamente N individuos para la siguiente generación.

Las diferencies fundamentales entre el método de Programación Evolutiva y la técnica de Algoritmos Genéticos son dos:

En primer lugar, no se impone restricción alguna sobre la representación. La aproximación típica del Algoritmo Genético implica la

codificación de las soluciones del problema en forma de una cadena de muestras representativas, el genoma. En la aproximación de la Programación Evolutiva la representación surge del problema.

En segundo lugar, la operación de mutación simplemente modifica aspectos de la solución según una distribución estadística que pondera como muy probables variaciones poco importantes en la descendencia y, como cada vez menos probables, variaciones considerables a medida que se vaya aproximando al óptimo máximo.

2.1.2.- Estrategias Evolutivas (EE)

Surgen en Alemania en la Universidad Técnica de Berlín en experimentos que utilizaban el túnel de viento del Instituto de Ingeniería de Flujos. Durante la búsqueda de las formas óptimas de los cuerpos en un flujo, lo que se reducía entonces a una experimentación laboriosa e intuitiva, se concibió la idea de adoptar un procedimiento estratégico. Sin embargo, los intentos con las estrategias de gradiente coordinado y simple fracasaron.

Entonces se les ocurrió probar con cambios aleatorios en los parámetros que definían la forma, siguiendo el ejemplo de las mutaciones naturales y nació así la EE. Se empezó a construir un experimentador automático, que funcionaría según las sencillas reglas de mutación y selección y se probó la eficiencia de estos nuevos métodos. Recibieron el apoyo financiero de la Sociedad de Investigación Alemana (DFG) que posibilitó el trabajo que se concluyó, de forma provisional, en 1974 con la tesis "Estrategia de evolución y la optimización numérica". De este modo, se crearon EE para resolver problemas de optimización técnica, y hasta hace muy poco éstas sólo se conocían entre la comunidad de la ingeniería civil, como una alternativa a las soluciones estándares.

2.1.3.- Algoritmos Genéticos

El algoritmo genético es un modelo de aprendizaje automático cuyo comportamiento deriva de un metáfora de los procesos de la evolución en la naturaleza. Su funcionamiento es mediante la creación, en una computadora, de una población de individuos representados por los cromosomas que son, en esencia, un conjunto de cadenas de caracteres

análogos a los cromosomas de 4 bases que vemos en nuestro ADN. Los individuos en la población se someten a un proceso de evolución.

La evolución no es un proceso intencionado ni dirigido, es decir, no existen datos para apoyar la afirmación que la evolución tiene como meta la producción de la humanidad. De hecho, los procesos de la naturaleza parecen reducirse a la competencia entre diferentes individuos por los recursos en el entorno. Algunos son mejores que otros y los mejores tienen una probabilidad más alta de sobrevivir y propagar su material genético.

En la naturaleza, vemos que la codificación para nuestra información genética (genoma) se realiza de forma que permita la reproducción sexual. Ésta, facilita la creación de una descendencia radicalmente diferente desde el punto de vista genético, pero que tiene el mismo sabor general (especie). Al nivel molecular, lo que ocurre, simplificando al máximo, es que un par de cromosomas se encuentran, intercambian trozos de información genética y se separan. Esto es la operación de recombinación, llamada generalmente "cruce" debido a la manera en la que el material genético cruza de un cromosoma a otro.

La operación de cruce se produce en un entorno en el que la selección del que va a emparejarse es una función de la aptitud del individuo, es decir, el grado de éxito que tiene el individuo al competir en su entorno.

Algunos algoritmos genéticos utilizan una función sencilla de la medida de aptitud para seleccionar individuos (de tipo probabilístico) que van a someterse a los operadores genéticos como el cruce o la reproducción (propagación de material genético sin cambios). Otras implementaciones utilizan un modelo en el que compiten determinados individuos que son seleccionados en un subgrupo de forma aleatoria y se elige el más apto. Los dos procesos que contribuyen más en la evolución son el cruce y la selección/reproducción basada en la aptitud.

También la mutación juega un papel en este proceso, aunque éste no es tan dominante como generalmente se cree en el proceso de la evolución. Hay que insistir sobre manera en que el algoritmo genético, como una simulación de un proceso genético, no es una "búsqueda aleatoria" de una solución a un problema (un individuo altamente apto). El algoritmo genético utiliza procesos estocásticos, pero el resultado no es en definitiva aleatorio [Holland, 1992].

En la práctica, se puede implementar este modelo genético de la computación con matrices de bits o caracteres que representan las cromosomas. Las operaciones sencillas de manipulación de bits permiten la implementación de las operaciones de cruce, mutación, etc. Aunque se ha realizado una cantidad considerable de investigaciones sobre las cadenas de longitud variable y otras estructuras, la mayor parte del trabajo con los algoritmos genéticos se centra en las cadenas de caracteres de longitud fija. Estos son los aspectos decisivos que caracterizan la programación genética en la que no se tiene una representación de longitud fija, ni existe por lo general una codificación del problema.

Cuando se implementa un algoritmo genético, se suele realizar de una manera que comprende el siguiente ciclo: evaluar la aptitud de todos los individuos en la población; crear una nueva población a través de las operaciones como el cruce, la reproducción en función de la aptitud y la mutación sobre los individuos cuya aptitud haya sido medida; desechar la vieja población e iterar con la nueva población (generación). No existe ninguna justificación teórica de este modelo de implementación pues no se aprecia este comportamiento puntual en las poblaciones en la naturaleza en su conjunto. La primera generación (generación 0) de este proceso opera sobre una población de individuos generada de forma aleatoria. A partir de ahí, las operaciones genéticas, conjuntamente con la medida de aptitud, tienen la función de mejorar la población.

La facultad de los algoritmos genéticos para concentrar su atención en las porciones más prometedoras de un espacio de soluciones es fruto directo de su capacidad para intercambiar cadenas portadoras de soluciones parciales. Al empezar, cada ristra de la población se evalúa con vistas a determinar el rendimiento de la estrategia codificada en ella, después, las que logran mayor puntuación se aparean entre sí: se alinean, se selecciona un punto al azar y se procede al intercambio de las subcadenas situadas a la izquierda del mismo. Se engendran así dos descendientes: uno que contiene los símbolos de la primera ristra hasta el punto de cruce más los de la segunda a partir de ese punto y, el otro formado por la hibridación complementaria. Los cromosomas biológicos se intercambian al encontrarse dos gametos para constituir un zigoto, por lo que el proceso de hibridación de algoritmos genéticos remeda de cerca el modelo biológico en que se inspira [Holland, 1992].

El algoritmo genético explota las regiones de más alto rendimiento del espacio de soluciones (las regiones "diana") porque las sucesivas generaciones de reproducción y cruzamiento generan un número creciente de ristras pertenecientes a ellas. El algoritmo favorece que las ristras más aptas asuman roles progenitores y, por esta razón, las cadenas superiores al promedio (que yacen en regiones diana) tendrán mayor descendencia en la generación siguiente.

La clave de esta conducta francamente sorprendente, reside en que cada ristra individual pertenece a todas las regiones en las cuales aparece uno cualquiera de sus bits. Así pues, un algoritmo genético que manipule una población de unos cuantos millares de cadenas está realmente tomando muestras de un número de regiones enormemente mayor. Tal paralelismo implícito (en el sentido de procesamiento en paralelo) proporciona al algoritmo genético su ventaja principal sobre otros procesos de resolución de problemas.

El paralelismo implícito del algoritmo genético le permite ensayar un gran número de regiones del espacio de búsqueda manipulando un número relativamente reducido de ristras. Resulta evidente que la evolución biológica no opera en el sentido de producir un único superindividuo, sino en el de generar especies bien adaptadas unas a otras, en interacción mutua. Del mismo modo, resulta posible utilizar el algoritmo genético, modificado, no para gobernar la evolución de meras reglas o estrategias individuales sino de "organismos" compuestos por multitud de reglas. En lugar de seleccionar aisladamente las reglas más idóneas, las presiones competitivas pueden llevar a la evolución de sistemas mayores cuyas facultades se encuentren codificadas en las ristras de bits que los componen.

2.1.4.- Sistemas Clasificadores (SC)

Surgen con la pretensión de extender el alcance los algoritmos genéticos y crear así un código genético capaz de representar la estructura de cualquier programa. El resultado fue lo que se denominó sistema clasificador, que consiste en un conjunto de reglas cada una de las cuales realiza determinadas acciones siempre que algún dato satisfaga sus condiciones. Condiciones y acciones están representadas por ristras o cadenas de bits correspondientes a la presencia o ausencia de

características específicas en las entradas o salidas de las reglas. Por cada característica que se halle presente la ristra contendrá un 1 en la posición adecuada y por cada una ausente un 0.

La forma de obtener por evolución un conjunto de reglas clasificadoras capaces de resolver un problema dado es la siguiente: se parte de una población de ristras aleatorias de unos y ceros y se evalúa cada una de ellas atendiendo a la calidad de su resultado. Según el problema de que se trate la medida de adecuación puede ser el beneficio comercial, el tanteo en un juego, la tasa de error o cualquiera de otros múltiples criterios. Las cadenas de mayor calidad se aparean; las inferiores perecen. Con el transcurso de las generaciones irán predominando las cadenas asociadas a soluciones cada vez más perfectas. Además, el proceso de apareamiento combina sin cesar estas ristras de nuevas formas, generando soluciones cada vez más refinadas. Los tipos de problemas que han resultado adecuados para esta técnica van desde el desarrollo de nuevas estrategias en la teoría de juegos hasta el diseño de sistemas mecánicos complejos.

2.1.5.- Programación Genética (PG)

La programación genética es la extensión del modelo genético de aprendizaje al espacio de los programas. Es decir, los objetos que constituyen la población no son cadenas de caracteres de longitud fija que codifican posibles soluciones al problema planteado, son programas que, cuando se ejecutan, "son" las soluciones candidatas para el problema. Estos programas se expresan en la programación genética como árboles de análisis gramatical, en vez de líneas de código. En la PG, la operación de cruce, se implementa al tomar subárboles seleccionados de forma aleatoria entre los individuos (selección según la aptitud) y su intercambio. Es importante señalar que la PG no suele emplear mutación como un operador genético.

2.1.6.- Utilidad de los AE

En principio, los AE pueden computar cualquier función computable, es decir, todo lo que puede realizar un ordenador digital normal. Los AE son especialmente mal adaptados a problemas cuyo modo de resolución ya es conocida, a no ser que se pretenda que estos Problemas sirvan como

programas de referencia. Los algoritmos de propósito especial, es decir, los algoritmos que tengan codificado un cierto grado de conocimiento del dominio del problema superarán normalmente a los AE en el rendimiento. Los AE deben utilizarse cuando no existe ninguna estrategia de resolución de problemas conocida y el dominio del problema es completo. Ahí es donde entran en juego los AE: encuentran soluciones de forma heurística en la ausencia de otras alternativas.

2.1.7.- Aplicaciones de AE que han tenido éxito:

"Timetabling": un tema que se ha abarcado con bastante éxito con los AG es la planificación de horarios. Una manifestación bastante común de este tipo de problema es la planificación de horarios de exámenes o de clases universitarias, etc. La naturaleza modular de la función de aptitud constituye la clave de la principal fortaleza potencial del uso de los AG para este tipo de cuestiones frente al uso de los métodos convencionales de programación por búsqueda y/o por restricciones. Otros dominios de aplicación son "Job-Shop Scheduling", "Ciencias de la Gestión" y "Game Playing".

2.1.8.- Computación Evolutiva y otras ciencias:

La computación evolutiva es un campo que interesa a investigadores de múltiples y dispares disciplinas. Así, según se muestra en la siguiente tabla, son múltiples los campos en los que la CE tiene aplicación:

Informáticos	Interesan el aprendizaje y los sistemas adaptativos. Todavía es necesario construir el hardware necesario par trabajar con los AE en grandes problemas del mundo real. La computación masivamente paralela será de gran ayuda.
Ingeniería	Problemas de optimización
Robótica	Para la construcción de robots móviles que intentan navegar en entornos inciertos.

Ciencia cognitiva	Pueden utilizar sistemas clasificadores para describir modelos de pensamiento y sistemas cognitivos.
Físicos	Usan el hardware inspirado en CE como la "Connection Machine" de Hillis para modelizar problemas del mundo real que incluyen miles de variables que deben manejarse en paralelo.
Biólogos	Maynard, Dawkins y Collins ha puesto de manifiesto el interés de la CE en lo concerniente a la investigación en la biología.
Filósofos y otros	

3.- Contribuciones de la Informática en la Biología

Las técnicas computacionales empleadas en biología abarcan una gama enorme. Entre otras, incluye, tal como se muestra en la siguiente tabla, las siguientes:

TÉCNICAS COMPUTACIONALES EMPLEADAS EN BIOLOGÍA	
Algoritmo A *	Formación inductiva de categorías
Lingüística computacional	Redes de neuronas artificiales
Modelización cualitativa	Reconocimiento jerárquico de patrones
Razonamiento basado en casos	Inferencia deductiva en PROLOG
"Connection Machine"	Procesamiento visual dirigido a modelos
Propagación de restricciones	Sistemas Expertos

Mantenimiento de bases de conocimiento	Codificación mínima en longitud
Bases de datos orientadas a objeto	Simulación
Inducción de gramáticas de contexto libre	Técnicas de adquisición del conocimiento

Actualmente, la computadora es de gran valor para los biólogos moleculares que investigan en genética para poder llevar a cabo:

- Traducción de secuencias
- Cálculos de la composición de bases
- Frecuencias de dinucleótidos de codones
- Análisis de mapas de restricción
- Predicción de regiones de DNA que codifican para proteína
- Comparación de nucleótidos mediante análisis de repeticiones y palíndromos
- Predicción de estructuras secundarias de RNA y proteínas
- Búsqueda en bases de datos.

3.1.- Proyecto Genoma Humano

El Proyecto Genoma es, en definitiva, un trabajo de cartografía en el que se dibujarán moléculas de DNA (ácido desoxirribonucleico), el compuesto que, enrollado sobre sí mismo formando un cromosoma, da origen a la vida. Para Nat Goodman, investigador principal del Instituto Whitehead de Boston, subsidiario del MIT, "la iniciativa para el estudio del genoma humano es una de las gigantescas aventuras científicas de la historia" [Frenkel, 1991].

Este proyecto abarca dos mapas distintos pero complementarios. El primero comprende la secuenciación completa de las 3.000 millones de bases. El otro, la ubicación precisa de los genes dentro de los cromosomas y su posterior descifrado para conocer la función exacta que realizan. Como cabe suponer, completar la lectura exhaustiva de un libro que, escrito sólo con las iniciales de los cuatro elementos químicos

primarios, ocuparía 200.000 páginas, representa el punto final de complicadas investigaciones en las cuales se hallan involucradas varias ramas de la ciencia.

Secuenciar el genoma humano permitirá definir con precisión sus productos, paso fundamental en la capacidad de analizar la contribución funcional de las variaciones de las secuencias. Al comparar secuencias humanas con las de otras especies se obtienen relaciones evolutivas de gran valor y ponen de manifiesto claves que revelan el lenguaje de control según el cual se consiguen los complejos patrones de expresión génica [Watson, 1990].

De este modo, secuenciar el genoma humano vendrá acompañado indefectiblemente por un paralelo incremento de la información de secuencias correspondientes a otras especies. Además, todos los aspectos del proceso de secuenciación están siendo actualmente automatizados con el fin de reducir los costes del proyecto, el de mayor presupuesto en la historia de la Biología. Estos costes resultan tan prohibitivos que el esfuerzo tendrá que ser compartido internacionalmente. Así, se tendrá que implementar una red telemática internacional de bases de datos de secuencias.

Para el desarrollo del proyecto HUGO se plantearon una serie de líneas u objetivos principales que se definen en los siguientes focos del proyecto:

- Establecer principalmente bases de datos conteniendo información sobre secuencias del DNA y los genes y nuevas funciones para identificar genes y otras informaciones conocidas.

- Crear mapas de cromosomas humanos consistentes en marcadores de DNA que científicamente permitan localizar rápidamente genes.

- Crear un depósito de material de investigación incluyendo conjuntos ordenados de fragmentos de DNA que representen completamente el DNA en los cromosomas humanos.

- Desarrollar nuevos instrumentos para analizar el DNA.

- Desarrollar nuevas formas de analizar el DNA incluyendo técnicas físicas y bioquímicas y métodos computacionales.

- Desarrollo de recursos similares en otros organismos internacionales que faciliten la investigación biomédica.

- Determinar la secuencia del DNA de una fracción larga del genoma humano y la de otros organismos.

En Estados Unidos son varios los organismos e instituciones implicados en este proyecto, así están: el National Institute of Health (NIH), el Departament of Energy (DOE), la National Science Foundation (NSF) y el Howard Hughes Medical Institute (HHMI). Reconociendo que este proyecto generará gran cantidad de datos biológicos que requerirán una gran iniciativa informática, se han destinado el 20% de los fondos del mismo a este esfuerzo. Esto es importante pues esta iniciativa podría incentivar el desarrollo de una tercera generación de bases de datos, de la misma manera que el comercio y la industria fomentaron el desarrollo de las bases de datos relacionales ya implantadas [Frenkel, 1991].

Se precisará moverse en procesos simbólicos y numéricos por lo que los requisitos informáticos incluyen el desarrollo de los mejores y más rápidos algoritmos a fin de comparar datos de secuencias y para predecir los plegamientos protéicos, lo que implica combinatoria, procesamiento paralelo y supercomputadoras. Asimismo, la IA y sus aplicaciones (redes de neuronas), estarán implicadas en la investigación para definir los lugares de unión entre los fragmentos de DNA [Frenkel, 1991].

3.2.- Bases de datos biológicas

Cuando en 1983, Doolittle y colaboradores, utilizaron una base de datos de una secuencia genética recién sintetizada para probar que un gen causante del cáncer era un pariente próximo del factor de crecimiento normal, los laboratorios de biología molecular en todo el mundo empezaron a instalar computadoras o acogerse a redes para realizar búsquedas en bases de datos biológicas. Desde entonces han surgido una variedad desconcertante de recursos informáticos para la biología. Estas bases de datos y otros recursos constituyen un servicio valioso no sólo para la comunidad biológica, sino también para el informático que busca información sobre este dominio.

Existe una base de datos que maneja estos recursos y que se mantiene en el National Laboratory de Los Alamos. Se llama LiMB y contiene las descripciones, los contactos y los métodos de acceso a más de 100 bases

de datos relacionadas con la biología molecular siendo, por tanto, una herramienta valiosa para localizar información [Lawton, 1989].

Otro fuente general de información sobre bases de datos es el National Center for Biotechnology Information (NCBI), que forma parte de la National Library of Medicine. El NCBI también proporciona una base de datos de secuencias genéticas desde comienzos de 1992.

Existen varias bases de datos de secuencias genéticas y que mantienen una coordinación cada vez más estrecha. Son el Genbank [Moore, 1990], la Nucleotide Sequence Database del European Molecular Biology Laboratory (EMBL) [Hamm, 1986] y la DNA Database de Japón (DDBJ) [Miyazawa, 1990].

La base de datos de secuencias proteínicas principal es el Protein Identification Resource (PIR) [George, 1986]. El NCBI también proporciona una base de datos en la que se encuentran combinación de secuencias proteínicas no redundantes.

Varias bases de datos contienen información sobre las estructuras tridimensionales de moléculas. Así, la Protein Data Bank (PDB), mantenida por el Brookhaven National Laboratory, contiene datos sobre la estructura de las proteínas, sobre todo a partir de datos cristalográficos. BioMagRes (BMR) es una base de datos obtenidos por NMR sobre proteínas, incluyendo estructuras tridimensionales, que mantiene la Universidad de Wisconsin, Madison [Ulrich, 1989]. Carbbank contiene información estructural para hidrocarbonatos complejos [Doubet, 1989].

El Online Registry File del Chemical Abstracts Service (CAS) es una base de datos comercial que contiene más de 10 millones de sustancias químicas. La Cambridge Structural Database es una base de datos que contiene información sobre pequeñas estructuras moleculares. Bases de datos de mapas genéticos (GMB), así como una base de datos de enfermedades humanas heredadas y sus características (OMIM) se mantienen en el Welch Medical Library en la Universidad John Hopkins. Existe una base de datos con información sobre los compuestos implicados en el metabolismo intermediario, llamada CompoundKB y disponible en el NCBI [Hunter, 1993].

Como muestra del tamaño de estas bases de datos biológicas, se comparan en la siguiente tabla las tres más importantes:

NOMBRE DE LA BASE DE DATOS	NÚMERO DE ENTRADAS	NÚMERO DE NUCLEÓTIDOS O AMINOÁCIDOS	ESPACIO DE DISCO REQUERIDO
GenBank 72.0	71280	92160761	280 MB
EMBL 31.0	72481	94390065	300 MB
PIR 32.0	40298	11831134	110 MB

Las base de datos de nucleótidos se duplican en tamaño aproximadamente cada 1,5 años.

Por último, uno de los recursos más importantes para un informático interesado en biología molecular es el sistema de tablón de anuncios bionet. Este tablón está disponible, a través de usenet, por correo electrónico.

4.- Bibliografía

- Creighton, T.E.: *Proteins, structures and molecular properties.* Freeman and Company. 1983.
- Doubet, S.; Bock, K.; Smith, D.; Albersheim, P. and Darvill, A.: *The Complex Carbohydrate Structure Datahase.* Trends in Biochemical Sciences, 14: 475. 1989.
- Frenkel, K.A.: *The Human Genome Project and Informatics.* Communications of the ACM. Vol. 34 (11): 41-51. Noviembre 1991.
- George, D.; Barker, W. and Hunt, L.: *The Protein Identification Resource.* Nucleic Acids Research, 14: 11-15. 1986.
- Goodenough, U.: *Genética.* Ediciones Omega S.A, Cap. 2 y 3, págs. 12-39. Barcelona. 1981.
- Guyton, F.: *Tratado de Fisiología Médica.* Editorial Interamericana. Mexico, 1984.
- Hamm, G. and Cameron, G.: *The EMBL Data Library.* Nucleic Acids Research, 14: 5-9. 1986.
- Holland J.H.: *Adaptation in Natural and Artificial Systems.* MIT Press. 1992.
- Hunter, L.: *Artificial Intelligence and Molecular Biology.* AAAI Press/MIT Press. Massachusetts-California, 42-44. 1992.
- Lawton, J.; Burks, C. and Martinez, F.: *Overview of the LiMB Database.* Nucleic Acids Researh 17: 5885-5899. 1989
- Lehninger, A.L.: *Bioquímica.* Ediciones Omega, S.A. Barcelona. Cap. 28-34. Parte 4. 1972.
- Miyazawa, S.: *DNA data Bank of Japan: Present status and future plants.* Computers and DNA, 7: 47-61. 1990.
- Moore, J.; Benton, D. And Burks, C.: *The Genbank Nucleic Acid Data Bank.* Focus II (4): 69-72. 1990.
- Ulrich, E.; Markley, J. and Kyogoyu, Y.: *Creation a Nuclear Magnetic Resonance Data Repository and Literature Base.* Protein Sequence and Data Analysis, 2 23-37. 1989.

- Watson, J.: *The Human Genome Projetc: Past, Present and Future.* Sciencie 248: 44-48. 1990.
- Watson, J.D. and Crick, F.H.C.: *Molecular Structure of Nucleic Acid. A structure for Deoxiribose Nucleic Acid.* Nature 171: 737-738. 1953.