GENERADOR DE GENES (GENGEN)

GABRIEL GARDUÑO SOTO Y HUGO PADILLA CHACÓN

DIVISIÓN DE ESTUDIOS DE POSGRADO DE LA
FACULTAD DE FILOSOFÍA Y LETRAS UNAM.
MÉXICO

Introducción

En el seno de la biología moderna permanece, casi intacto, un núcleo de problemas fundamentales, a los cuales no se les ha encontrado una cabal solución. Problemas tan elementales como la definición misma del ser vivo, la dilucidación de complejas relaciones metabólicas, el mecanismo de acción de la autocatálisis del ARN; el problema del origen de la vida y de la evolución molecular abiótica son tan sólo algunos de los problemas que mantienen ocupados a los pensadores biológicos de este siglo.

Actualmente se han elaborado esquemas cognoscitivos provenientes de distintos horizontes epistemológicos para aproximarse conceptualmente hacia la solución de dichos problemas, pues la envergadura de éstos ha exigido a la disciplina biológica un replanteamiento de sus presupuestos más básicos —ya sean de naturaleza ontológica, lógica o metodológica—, tal como aconteciera previamente con la ciencia física, a principios del siglo. Y al igual que en la física, hoy en día en la biología, el cultivo de las disciplinas y técnicas computacionales ocupa un lugar destacado como adyuvante de este colosal esfuerzo que despliega el hombre como especie por allanar el camino hacia la comprensión del mundo viviente.

En este marco de circunstancias, vemos cómo diversas visiones científicas —estocásticas o deterministas— encuentran en la computación la herramienta de trabajo sin la cual, no sería posible, ni siquiera aventurar hipótesis que para poder ser planteadas y exploradas, requieren de una gran cantidad de cálculos en tiempos finitos y accesibles.

Uno de estos casos es el que nos proponemos tratar en esta reunión. A saber: el problema que representa nuestra actual incapacidad para explicar la organización secuencial de las bases púricas (Adenina y Guanina) y pirimídicas (Citosina, Timina -Uracil en ARN-) en la formación de una hebra de ácido nucleico (ADN o ARN).

Antecedentes de la problemática

Jacques Monod plantea en su libro, *El Azar y la Necesidad* (1970), la aparentemente insalvable dificultad para predecir las secuencias genómicas o para encontrar un orden en la sucesión de ellas; sin embargo, no descarta la posibilidad de que en el futuro la razón o conjunto de reglas que gobiernan las secuencias genómicas, pudieran ser encontradas.

Así pues, henos aquí veintidos años después de la publicación de *El Azar y la Necesidad*, y si bien aún no se ha encontrado una solución completa al problema que inquietaba a Monod, es necesario reconocer que se han dado algunos pasos adelante, pues Monod plantea en su libro como hipótesis la existencia de una invariancia de las estructuras genéticas, responsables de las secuencias polipeptídicas, que sin embargo se encontrarían organizadas puramente al azar. A continuación citaremos a Jacques Monod:

> La primera secuencia completa de una proteína globular fue descrita en 1952 por Sanger. Esto fue a la vez una revelación y una decepción. En esta secuencia en la que se sabía definir la estructura, o sea las propiedades electivas de una proteína funcional (la insulina), ninguna regularidad, ninguna singularidad, ninguna restricción se revelaba. Aunque, sin embargo, se puede esperar que, a medida que se acumulen documentos de este tipo, algunas leyes generales de ensamblaje, así como ciertas correlaciones funcionales, se harán más claras… De estas secuencias, y de su comparación sistemática ayudada por los modernos medios de análisis y de cálculo, se puede hoy deducir la ley general: la del azar.
> Para ser más precisos: estas estructuras están «al azar» en el sentido que, conociendo exactamente el orden de 199 residuos en una proteína que comprende 200, es imposible formular ninguna regla, teórica o empírica, que permita prever la naturaleza del único residuo no identificado aún por el análisis.[1]

Como podemos ver en la cita, el azar al cual se refiere Monod, no es un azar intrínseco, acerca del cual se tuviera una prueba formal de su aleatoriedad; el azar de Monod es más bien un azar provisorio, un concepto-herramienta, para tratar la complejidad de las estructuras secuenciales que aún no podemos interpretar, en el caso de que efectivamente, existan estructuras en dichas secuencias.

Regresemos al texto original de Monod:

> Decir de la secuencia de los aminoácidos en un polipéptido que está «al azar», no agrega nada, hay que insistir en ello, a una declaración de ignorancia; sino que expresa una constatación de hecho: por ejemplo, frecuencia media con la cual el residuo está seguido de uno determinado en los polipéptidos es igual al producto de las frecuencias medias de cada uno de los residuos en las proteínas en general.[2]

Cabe mencionar que el resultado anterior fue obtenido analizando centenares de secuencias con una cantidad no muy grande de nucleótidos. Actualmente, con los métodos de análisis bioquímico como los secuenciadores automáticos y las computadoras modernas es posible estudiar y analizar miles y hasta millones de dichas secuencias. En este universo expandido de análisis han comenzado a surgir algunos indicios acerca de la posibilidad de que «el azar» en las secuencias genómicas, conozca ciertas estructuras. Si bien sigue siendo cierto que aún no podemos interpretarlas.

De hecho, en la década de los ochenta el análisis computarizado de ingentes colecciones de datos genómicos comenzó a arrojar alguna luz sobre la posible existencia de regularidades en las cadenas genéticas. Ruth Nussinov (1987) revisa los reportes existentes sobre el tema y nos dice:

> Más recientemente ha sido mostrado que las frecuencias de ocurrencia de dinucleótidos no fluctúan aleatoriamente en secuencias de ADN de procariotes ni de eucariotes… Algunos dinucleótidos se encuentran presentes más que otros, por ejemplo AA > AT, GC > CA y TG > TA en secuencias procarióticas, o TG > GA, CT > AC y CA > TC en secuencias eucarióticas. También es posible discernir jerarquías, *e.g*. GC > AT > GT > TA en la mayoría de las secuencias procarióticas probadas, con aproximadamente 122, 000 nucleótidos o GC > GC > GT > TA > CG en la mayor parte de las 256 secuencias eucarióticas analizadas (con exclusión de las mitocondriales) totalizando aproximadamente 290,000 nucleótidos.[3]

Recientemente se han analizado reportes sobre las frecuencias de ocurrencia de codones (tripletes de bases púricas y pirimídicas) en las cadenas genómicas:

> Entonces, los codones deberían estar presentes en aproximadamente iguales frecuencias y los aminoácidos en las proteínas a frecuencias aproximadamente proporcionales al número de codones que los codifican. Se conoce ahora que esto no sucede.[4]

Hasta la fecha no se ha publicado ningún reporte fidedigno que haya resuelto el problema biocibernético planteado por Jacques Monod en 1970.

El “Generador de Genes”

En este marco de referencias, los autores de este trabajo, presuponen, como hipótesis de trabajo, la posibilidad de que las secuencias genómicas —y por ende las secuencias polipeptídicas— posean una estructura —o familias de estructuras— determinadas.

En el desarrollo del trabajo presentado, se revisaron distintos enfoques teóricos y computacionales —modelos de tipo biocibernético,[5,6,7] estadístico, pseudo-azaroso,[8] fractal,[9] markoviano, de autómatas[10,11] y autómatas celulares,[12,13] también se revisaron estudios sobre las necesidades termodinámicas que prevalecen en el

microambiente intermolecular del acoplamiento secuencial de las mencionadas bases púricas y pirimídicas. (Estudios estos últimos, que rebasaban los propósitos y recursos de los autores)—, que pudieran servir de apoyo para lograr mediante simulación computarizada la creación original de cadenas de nucleótidos bajo una razón de construcción lógico-matemática, cadenas que una vez creadas estarían destinadas a la búsqueda de sus homologías con secuencias genómicas reales.

Encontrando que los enfoques más recientes desarrollaban, salvo algunas excepciones, a lo sumo juegos matemáticos formales —como el 'Juego de la Vida'—,[14] los autores decidieron, a su vez realizar un enfoque propio, basado en la teoría de números y en la lógica clásica. Ya que nuestro propósito inicial fue ilustrar, mediante un ejemplo biológico, otras aplicaciones de la lógica clásica y sus posibilidades. En rigor, nuestro enfoque apunta hacia la dilucidación de las razones, que gobiernan la concatenación de las secuencias genómicas, que presuponemos expresables a través de un enfoque lógico, hipótesis atrevida, con miras a aplicar un modelo similar a las estructuras estéricas reales, dilucidadas o predichas, en el laboratorio químico o biomatemático, respectivamente —campo aún muy reciente—.

Así pues, en nuestro enfoque, hemos desarrollado una visión lógica del fenómeno, donde se simula una propuesta arbitraria que presupone en cada base púrica y pirimídica, dos sitios activos, sin ningún comportamiento de cooperación, disponibles para la interacción de concatenación, que se suponen, también arbitrariamente, como sitios que se comportan como una aplicación de los operadores lógicos tradicionales, para producir teóricamente la siguiente posición nucleotídica en la cadena. (En modo muy similar a la construcción de la serie de Fibonacci).

De este modo el programa concreto quedó organizado de la siguiente manera:

a) El programa lo constituyen varios subprogramas algorítmicos —algunos algoritmos son de orden 5—, posee un generador de números asociados mediante un algoritmo (no-azaroso y formalmente indecidible) y aplica a dichos números los operadores lógicos tradicionales y combinaciones de ellos. El proceso anterior vierte sus resultados de acuerdo con el código genético, para ir produciendo la cadena de secuencias genómicas. En cada caso, el programa parte de condiciones iniciales definidas por el usuario; a saber:
b) Elección de una asignación numérica para cada **base púrica o pirimídica**. (8 Opciones).[15]
c) Elección del **codón inicial** (numérico) correspondiente al iniciador bioquímico **Metionina**.
d) Elección del **codón secundario** (numérico) entre cualquiera de los 64 posibles tripletes del código genético.
e) Elección de dos números de "arranque" que funcionan como **semillero** para el generador numérico.
f) Elección de las condiciones de visualización de la secuencia genómica generada: 1) En código internacional de literales para cada aminoácido, 2) En abreviatura tradicional para aminoácidos, 3) En una secuencia propiamente genómica de **tripletes** de bases púricas y pirimídicas 4) En una concatenación numérica que representa la salida del programa, para fines de estudio.
g) Elección de lectura y desplegado para los *Stops* **Ambar**, **Ocre** y **Ópalo**.

Dada la longitud y profusión de los resultados, éstos se visualizan en pantalla y se registran en disco duro o flexible, para su análisis posterior.

El programa ofrece grandes posibilidades de "experimentación teórica" y de estudio de condiciones de concatenación genómica al usuario ya que sus posibles condiciones inciales son superiores a **treinta y dos mil doscientos millones de combinaciones**.

Resultados

El programa puede generar cadenas genómicas de longitud variable, dependiendo de las condiciones de arraque, en teoría podría generar cadenas de longitud ilimitada, dependiendo del tiempo de proceso disponible. La operación se lleva a cabo en computadoras PC con un mínimo de 640 Kb de memoria RAM, y es completamente portable a cualquier sistema de cómputo independientemente del lenguaje de programación que se utilice, dado que la efectividad del sistema reside principalmente en los algoritmos básicos de operación autónoma y no en aplicaciones dependientes del lenguaje usado ni de la versión utilizada.

Por otra parte, en su operación, el programa muestra comportamiento de auto-organización y actualmente muestra una tendencia hacia la estabilización de las secuencias producidas. Se piensa introducir posteriormente rutinas de autoregulación, para aumentar la longitud de las cadenas producidas.

Actualmente se analizan algunas salidas interesantes del programa investigando homologías con cadenas de nucleótidos reales, a través de consulta directa con los bancos de datos del proyecto Genoma Humano y de GenBank.

Un producto secundario del programa lo constituye la utilidad de almacenamiento compactado de las cadenas artificiales, que compartan homologías con cadenas reales, en los sistemas de cómputo, con ahorro de memoria y tiempo de transmisión por vías de comunicación electrónica.

Bibliografía

[1] Jacques Monod. *El Azar y la Necesidad*. *Ensayo sobre la filosofía natural de la biología moderna*. Cuadernos Infimos 100. Tusquets, Barcelona, 1981. (Edición original en francés de Editions de Seuil, Paris, 1970).

[2] Ibid.

[3] Ruth Nussinov. “Theoretical Molecular Biology: Prospectives and Perspectives”. *J. Theor. Biol*. 125: 219-235, 1987.

[4] Curnow. R.N. “The Use of Markov Chain Models in Studying the Evolution of the Proteins”. *J. Theor. Biol*. 134: 51-57, 1988.

[5] Michael Conrad. “Molecular Computing as a Link between Biological and Physical Theory”. *J. Theor. Biol*. 98: 239-252, 1982.

[6] Motoyosi Sugita. “Functional Analysis of Chemical Systems *in vivo* using a Logical Circuit Equivalent”. *J. Theor. Biol*. 1: 415-430, 1961.

[7] Motoyosi Sugita. “Functional Analysis of Chemical Systems *in vivo* using a Logical Circuit Equivalent. II. The Idea of a Molecular Automaton”. *J. Theor. Biol*. 4: 179-192, 1963.

[8] Kauffman S.A. “Metabolic Stability and Epigenesis in Randomly Constructed Genetic Nets”. *J. Theor. Biol*. 22: 437-467, 1969.

[9] David L. Finkel. “Fractal Displays of Genomic DNA. I. Eco RI Fractal Lattice of Buffalo Rat”. *Int. Jour. of Quantum Chemistry*. XXXVI: 575-586, 1989.

[10] Narendra S. Goel *et al*.: “Movable Finite Automata (MFA) Models for Biological Systems. II. Protein Biosynthesis”. J. Theor. Biol. 134: 9-49, 1988.

[11] René Thomas. “Regulatory Networks Seen as Asynchronous Automata: A Logical Description. — Mini Review”. *J. Theor. Biol.* 153: 1-23, 1991.

[12] Stauffer D. “Computer Simulations of Cellular Atomata”. *J. Phys. A: Math. Gen.* 24: 909-927, 1991.

[13] Wolfram S. “Universality and Complexity in Cellular Automata”. *Physica* 10D: 1-35, 1984.

[14] Bagnoli F., Rechtman R., Ruffo S. “Some Facts about Life”. *Physica A* 171: 249-264, 1991.

[15] La asignación numérica inicial y las transformaciones numéricas ulteriores fueron diseñadas por los autores para las necesidades del modelo. El lector interesado puede referirse a métodos similares en: Sepúlveda A. *et al*. “Storage and Retrieval of Biomolecule Sequences”. *J. Theor. Biol*. 103: 331-332, 1983. Johnson F.Y., *et al*. “Prime Numbers and the Amino Acid Code: Analogy in Coding Properties”. *J. Theor. Biol*. 151: 333-341, 1991.