INTELIGENCIA ARTIFICIAL GENERATIVA (GENAI) PARA DISEÑAR GENOMAS DE ORGANISMOS DIGITALES

Francisco J. Borrallo-Vázquez y Miguel A. Fortuna
Laboratorio de Biología Computacional
Estación Biológica de Doñana (EBD), CSIC, Sevilla

La Inteligencia Artificial Generativa (GenAI por sus siglas en inglés) es una rama de la Inteligencia artificial (IA) que se enfoca en la creación de nuevos contenidos, como imágenes, texto, música y otros datos, a partir de patrones y datos preexistentes. En lugar de simplemente analizar o clasificar información, los modelos de GenAI, como ChatGPT, aprenden a replicar y crear datos a partir de los que han usado durante su entrenamiento. Este tipo de algoritmos es particularmente útil en áreas creativas y de diseño, como la creación de arte digital, la generación de contenido en videojuegos, la composición musical y la simulación de conversaciones humanas. Pero además, la GenAI también tiene aplicaciones en ámbitos científicos, como el diseño de moléculas activas novedosas en farmacología, la predicción de patrones de coexistencia entre especies en comunidades ecológicas, y la generación de secuencias genómicas en la investigación biomédica.

La genómica sintética busca diseñar y construir genomas completos para encontrar los principios fundamentales responsables de la función del genoma. Los avances recientes en ensamblaje, edición, y síntesis de ADN junto con las innovaciones computacionales en la GenAI han impulsado este campo. El objetivo de la investigación que llevamos a cabo con genomas digitales es, precisamente, guiar el diseño y construcción de genomas sintéticos de organismos naturales. Sin embargo, trasladar lo aprendido *in silico* a los proyectos de genómica sintética sigue siendo un proceso altamente complejo. Nuestros esfuerzos en la aplicación de la GenAI a genomas de organismos digitales se centran en optimizar este proceso para intentar, en una segunda fase, trasladar lo aprendido al diseño, construcción, entrega, ajuste y aplicación de genomas sintéticos de organismos naturales.

Evolución digital

La evolución digital es una forma de computación evolutiva en la que programas informáticos autorreplicantes —organismos digitales— mutan y evolucionan dentro de un entorno computacional definido por el usuario. **Avida** es la plataforma de *software* más

utilizada para la investigación en evolución digital y se ha consolidado como el nexo de unión entre la simplicidad y abstracción de los modelos matemáticos, por un lado, y la complejidad y realismo de los experimentos en el laboratorio y en condiciones naturales, por el otro. **Avida** cumple con los tres requisitos esenciales para que tenga lugar el proceso evolutivo: replicación, variación heredable y eficacia biológica diferencial —fitnes-. Este último requisito emerge de la competencia por los recursos limitados de espacio en memoria (RAM) —donde "viven" los organismos digitales— y tiempo de la unidad central de procesamiento (CPU) que usan los organismos digitales para replicarse. Un organismo digital en Avida consiste en una secuencia de instrucciones —su genoma— y una CPU virtual que ejecuta esas instrucciones.

El espacio de secuencias que pueden codificar organismos digitales con genomas de longitud —número de instrucciones— L extraído de un alfabeto de instrucciones disponible A, comprende A^L genomas diferentes. Si consideramos genomas con L = 100 instrucciones tomadas de un alfabeto de A = 26 instrucciones (el lenguaje genético de Avida; Figura 1), el espacio de secuencias es enorme: $26^{100} = 3{,}14 \times 10^{141}$. Cualquier genoma en este espacio codificará un organismo viable si éste es capaz de autorreplicarse. Encontrar genomas viables en este gigantesco espacio mediante la generación de secuencias aleatorias es como buscar una aguja en un pajar.
El coste computacional es muy alto. Por ejemplo, se necesitan más de un millón de secuencias aleatorias para encontrar un solo genoma viable de 100 instrucciones. Por tanto, necesitamos métodos mucho más eficientes en la búsqueda de genomas que codifiquen organismos viables que puedan usarse como ancestros—*wild type*— a partir de los cuales iniciamos nuestros experimentos evolutivos destinados a desvelar los mecanismos responsables de la biodiversidad del planeta. Y es en este punto donde la GenAI entra en juego.

Redes generativas antagónicas

Las Redes Generativas Antagónicas (GANs por sus siglas en inglés) son una clase de algoritmos de GenAI que imitan la "carreras de armamentos" evolutivas que tienen lugar, por ejemplo, entre los depredadores y sus presas (donde el depredador evoluciona para ser más eficiente en la búsqueda y captura de su presa, y su presa evoluciona para evadir más fácilmente a su depredador). Una GAN consiste en dos

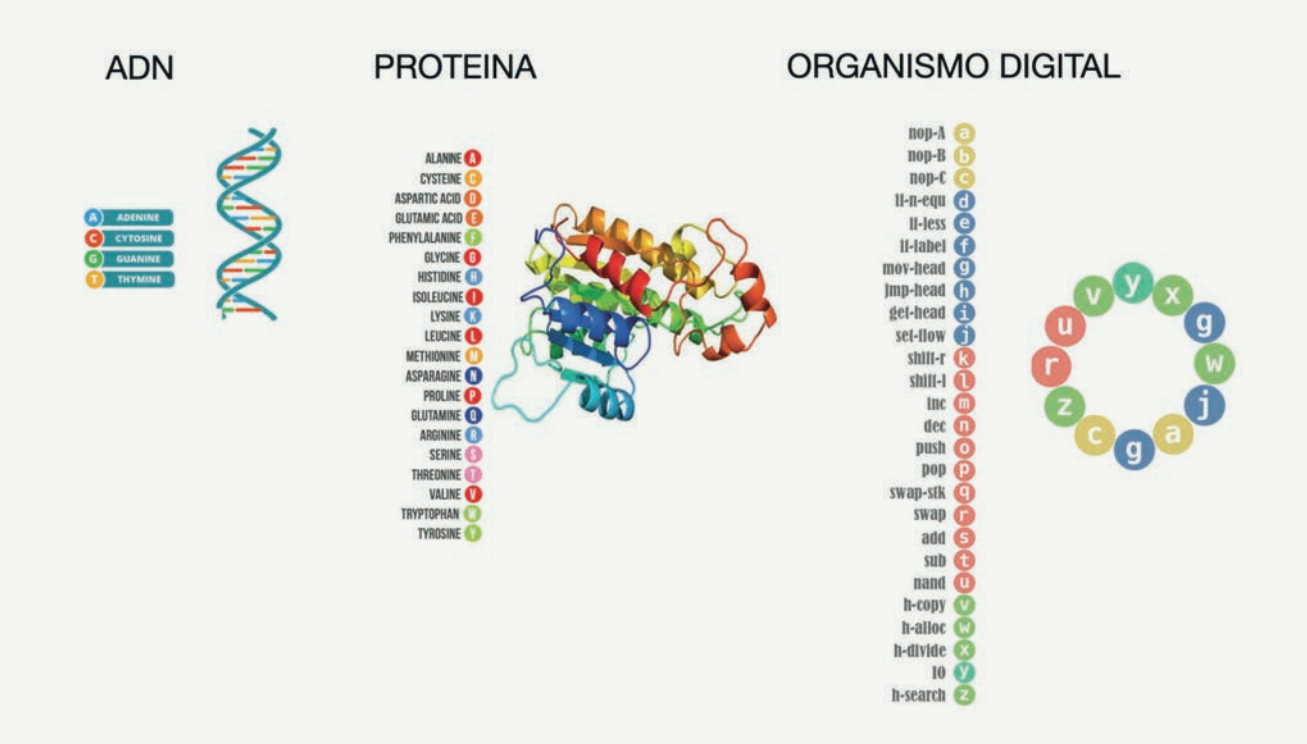

Figura 1
El lenguaje genético del genoma de los organismos digitales está formado por 26 códigos de instrucciones. Es decir, a diferencia de los cuatro nucleótidos que constituyen el ADN del genoma de los organismos naturales o de los 20 aminoácidos a partir de los cuales se forman las proteínas, cada posición del genoma de un organismo digital puede estar ocupada por una de las 26 "letras" posibles.

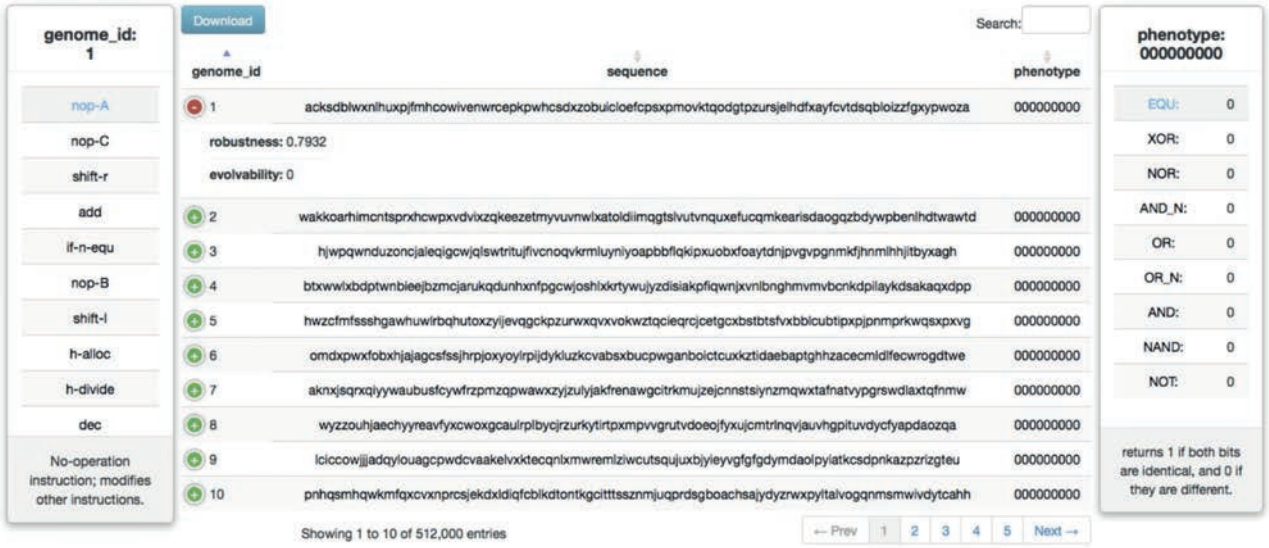

Figura 2
avidaDB es una base de datos que contiene más de un millón de genomas de organismos digitales disponibles para ser usados en experimentos evolutivos. Con el fin de facilitar el acceso a avidaDB hemos desarrollado una librería para el lenguaje de programación R (avidaR) que permite hacer descargas y búsquedas selectivas de genomas.

redes neuronales artificiales con múltiples capas que se entrenan de manera concurrente: una red generativa y una red discriminativa. Dentro de nuestro propósito de encontrar genomas que codifiquen organismos digitales viables, la red generativa (generador) usa ruido aleatorio gaussiano para producir genomas tan reales (es decir, genomas que codifican organismos viables) como sea posible, de modo que se entrena para generar genomas falsos (es decir, que no codifican organismos viables). La red discriminativa (discriminador) recibe genomas aleatorios del generador y genomas reales de la base de datos avidaDB (Figura 2), y decide si el genoma es real o no. Tanto la red generativa como la discriminativa se entrenan a la vez, jugando una contra la otra para maximizar sus objetivos: el del generador es proporcionar genomas aleatorios lo suficientemente reales como para engañar al discriminador (que los clasificaría como reales aunque no lo sean), y el del discriminador es obligar al generador a producir genomas ni tan reales ni tan aleatorios como para clasificarlos acertadamente como reales o falsos, respectivamente. Cuando este aprendizaje recíproco se equilibra, el generador proporciona genomas aleatorios con alta probabilidad de que codifiquen organismos digitales viables (Figura 3).

Una de las ventajas de usar esta aproximación computacional como gemelo digital del diseño de genomas de organismos naturales es que podemos validar el éxito del entrenamiento de las GANs inmediatamente y sin coste alguno. Es decir, a diferencia del diseño de secuencias genómicas mejoradas para la industria y la biomedicina que requieren posteriormente su síntesis química para corroborar que se ha implementado con éxito la funcionalidad biológica deseada, la comprobación de la viabilidad de los genomas de organismos digitales generados con las GANs se comprueba directamente en Avida. Sólo necesitamos ejecutar el código del organismo digital que contiene el genoma que proporciona el generador de la GAN y, en cuestión de milisegundos, observar si es o no capaz de replicarse. Y esta validación retroalimenta a su vez el sistema de entrenamiento de la GAN porque los genomas generados que codifiquen organismos viables formarán parte de la base de datos de genomas reales —avidaDB— con la que se entrenan tanto el generador como el discriminador.

No sólo evolución sino también coevolución digital

En Avida, los organismos digitales necesitan consumir recursos del entorno computacional para replicarse, de forma análoga al consumo de nutrientes por parte de las bacterias. Los recursos computacionales definidos en Avida se consumen sólo si el organismo es capaz de realizar funciones matemáticas sencillas con números binarios mientras ejecuta las instrucciones que constituyen su genoma. Y de manera análoga a como la bacteria *E. coli* metaboliza glucosa o citrato como fuente de carbono, un organismo digital puede consumir el recurso asociado a realizar una suma (operación booleana *OR*) o una multiplicación (operación booleana *AND*). Esta

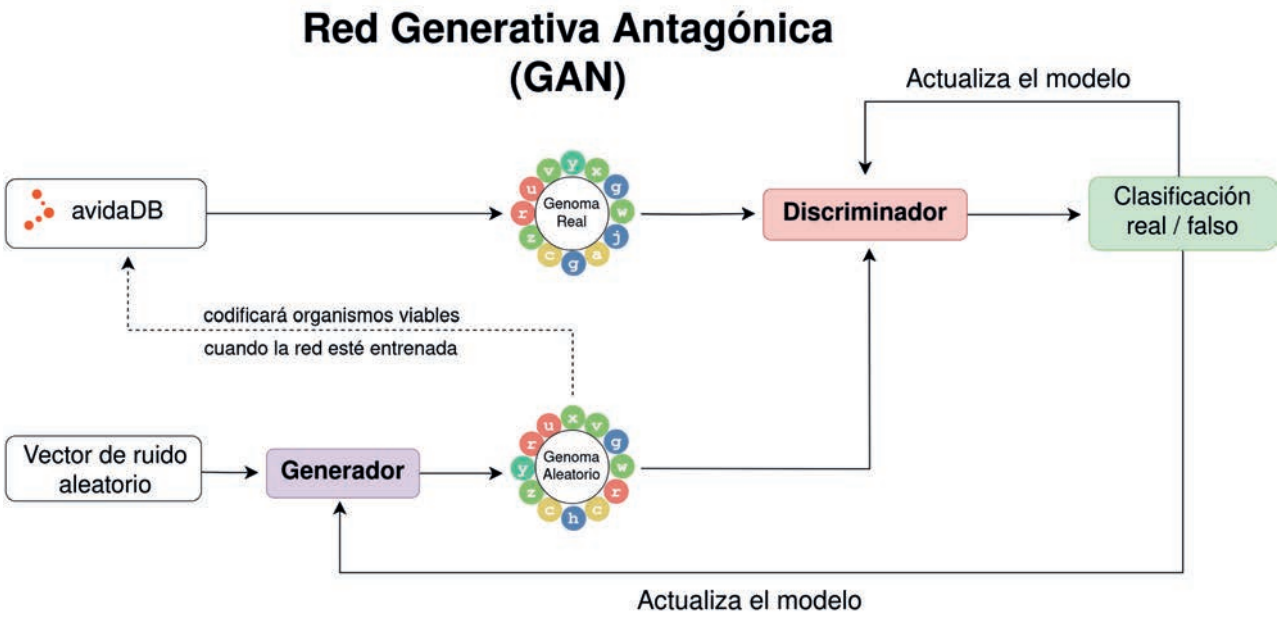

Figura 3

Una Red Generativa Antagónica (GAN) para el diseño de genomas consta de dos redes neuronales que se entrenan simultáneamente: una red generativa (generador), que usa ruido aleatorio para producir genomas aleatorios y se entrena para generar genomas falsos pero no tan aleatorios, y una red discriminativa (discriminador), que recibe genomas del generador y de la base de datos avidaDB, clasificándolos como reales o falsos. Ambas redes se entrenan una contra la otra para mejorar su precisión. Al equilibrarse este proceso, el generador produce genomas con alta probabilidad de codificar organismos digitales viables que, de serlo, pasarán a formar parte de la base de datos.

capacidad que tiene un organismo digital de realizar operaciones booleanas para consumir recursos define su fenotipo o funcionalidad.

Esta propiedad funcional de los organismos digitales ha permitido implementar interacciones hospedador-parásito, en donde algunos organismos —parásitos— son capaces de "robar" la energía que necesitan (ciclos de CPU) para ejecutar las instrucciones de sus genomas a otros organismos digitales —sus hospedadores. Un parásito podrá infectar a un hospedador si realiza el menos una de las funciones booleanas que lleva a cabo éste. Es el equivalente digital al acoplamiento entre las fibras de la cola de un virus bacteriófago y los receptores de membrana de la bacteria. La presión selectiva que imponen los parásitos beneficiará a aquellos hospedadores que acumulen mutaciones en el genoma que cambien la función booleana que realizan —el receptor de membrana bacteriano. De esta forma el hospedador podría escapar de los parásitos. Por tanto, si pensamos en el creciente interés que ha adquirido el uso de virus bacteriófagos como agentes terapéuticos en la lucha contra las infecciones bacterianas resistentes a los antibióticos, se requiere un método de generación de secuencias genómicas que posean la funcionalidad necesaria para infectar variantes genómicas que puedan emerger durante la evolución bacteriana (Figura 4).

De los genomas de organismos digitales a los genomas de organismos naturales

El objetivo fundamental del uso de gemelos digitales, como Avida, es complementar los estudios llevados a cabo en el laboratorio. El paso crucial es comprobar si lo que aprendemos estudiando organismos digitales en un ordenador es útil para entender los mecanismos y procesos que suceden en los organismos naturales. Y en este punto se puede ver el vaso medio vacío o medio lleno, es decir, fijarse únicamente en las diferencias entre organismos digitales y naturales (que las hay) o profundizar en las características comunes (que también las hay, y muchas). Para nosotros, el proceso evolutivo es independiente del substrato material que contiene y transmite la información, sea la química molecular basada en el carbono o el estado de los electrones en un semiconductor. Lo verdaderamente importante es que haya entidades replicantes (células o programas informáticos) que contengan información heredable (genoma) y que las variaciones en la información

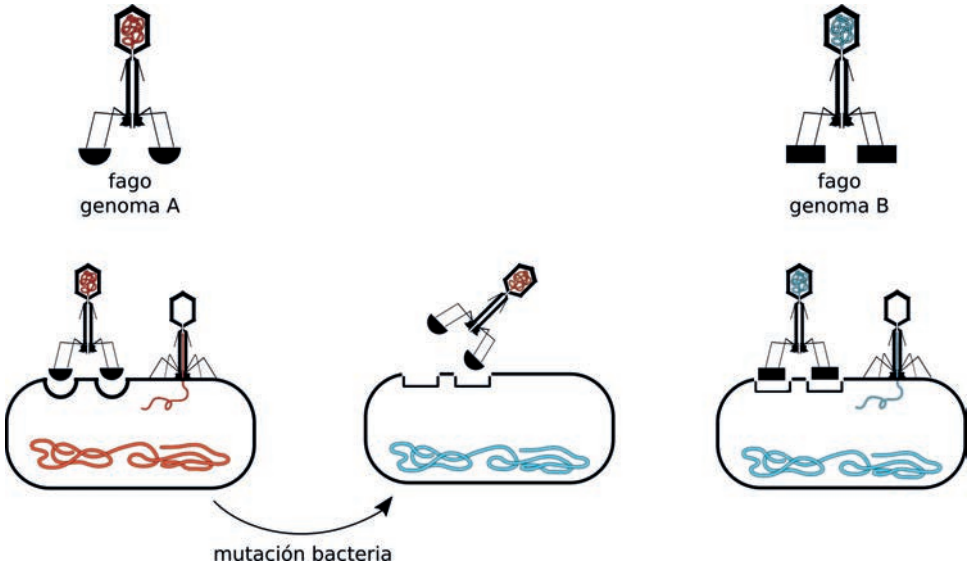

Figura 4
Un parásito digital podrá infectar a un hospedador digital si realiza el menos una de las funciones booleanas que lleva a cabo el hospedador. Es el equivalente digital al acoplamiento entre las fibras de la cola de un virus bacteriófago (fago) y los receptores de membrana de sus hospedadores bacterianos. Una vez se produce ese acoplamiento, el parásito infectará a su hospedador. Este proceso impone una presión selectiva que beneficiará a aquellos hospedadores que adquieran mutaciones que les permitan realizar nuevas funciones booleanas, de manera análoga a cambios en los receptores de membrana de las bacterias. De esta forma, el hospedador podrá escapar de los parásitos. Por tanto, se require un método de generación de genomas de parásitos digitales que posean la funcionalidad necesaria para infectar a la variante genética específica que pueda emerger durante la evolución del hospedador.

que contienen les otorguen mayor o menor eficacia en la transmisión de esa información durante su proceso de replicación.

¿Hasta qué punto este enfoque digital puede ayudar a diseñar nuevos genomas de organismos naturales? Si nos centramos en el uso de fagos como agentes terapéuticos, podemos entrenar las GANs con los genomas de los fagos y los de sus hospedadores bacterianos para ayudar a los investigadores a identificar nuevos hospedadores susceptibles de ser infectados. De hecho, el grupo de Jim Collins en el MIT ha desarrollado recientemente una herramienta llamada BioAutoMATED, que utiliza modelos de lenguaje entrenados en secuencias genómicas biológicas para interpretar y diseñar genomas que posean una funcionalidad específica, análoga a las funciones booleanas que son capaces de llevar a cabo los parásitos digitales en Avida.

Para leer más

Eugene L, *et al.* "Relevant applications of generative adversarial networks in drug design and discovery: molecular *de novo design*, dimensionality reduction, and *de novo* peptide and protein design". *Molecules* 25 (2020) 3250. https://doi.org/10.3390/molecules25143250

Fortuna MA, *et al.* "The genotype-phenotype map of an evolving digital organism". *PLoS Computational Biology* 13 (2017) e1005414. https://doi.org/10.1371/journal.pcbi.1005414

James JS, *et al.* "The design and engineering of synthetic genomes". *Nature Review Genetics* (2024) https://doi.org/10.1038/s41576-024-00786-y

Valeri JA, *et al.* "BioAutoMATED: an end-to-end automated machine learning tool for explanation and design of biological sequences". *Cell Systems* 14 (2023) 525-542. https://doi.org/10.1016/j.cels.2023.05.007

Yelmen B, *et al.* "Creating artificial human genomes using generative neural networks". *PLoS Genetics* 17 (2021) e1009303. https://doi.org/10.1371/journal.pgen.1009303