CSIC
Consejo Superior de Investigaciones Científicas

Nota de prensa

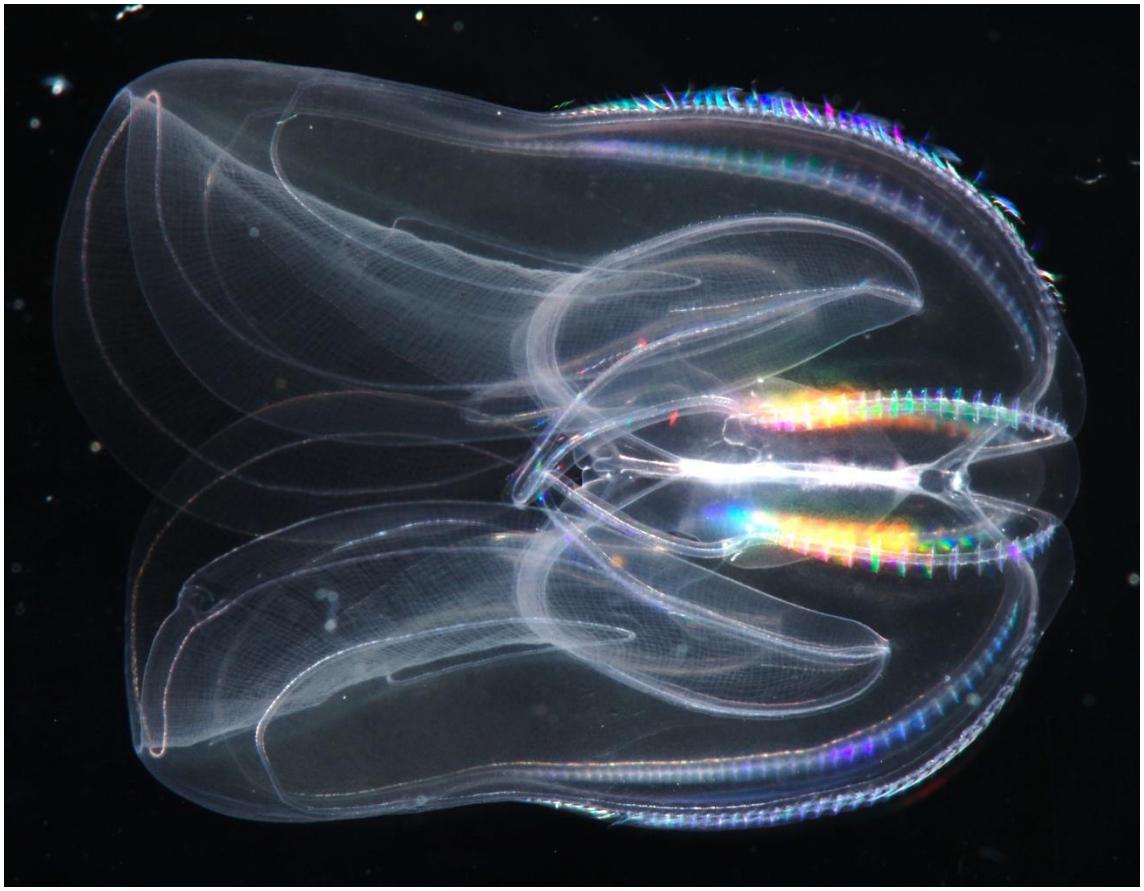
CSIC comunicación
Tel.: 91 568 14 77
comunicacion@csic.es
www.csic.es

Barcelona / Sevilla, martes 4 de noviembre de 2025

Una herramienta basada en IA es capaz de predecir funciones desconocidas de cualquier proteína

- **Un estudio del IBE (CSIC-UPF) y el CABD-CSIC ha revelado la función de 24 millones de genes codificantes para proteínas**

- **La nueva herramienta, con potencial en biología evolutiva, arroja luz sobre el "proteoma oscuro", el cual comprende la mitad de las proteínas cuya función se desconoce**



Invertebrado ctenóforo 'Mnemiopsis leidyi'. /Bruno C. Vellutini / Wikimedia Commons

Un equipo liderado por **Rosa Fernández**, del Instituto de Biología Evolutiva (IBE), un centro mixto del Consejo Superior de Investigaciones Científicas (CSIC) y la Universidad

Pompeu Fabra (UPF), y **Ana Rojas**, del Centro Andaluz de Biología del Desarrollo (CABD), un centro mixto del CSIC, la Junta de Andalucía y la Universidad Pablo de Olavide (UPO), ha desarrollado una herramienta basada en inteligencia artificial (IA) capaz de predecir la función desconocida de las proteínas a partir de secuencias genómicas sin una referencia previa, mediante la aplicación de modelos de lenguaje. En cuestión de horas y sin necesidad de un entrenamiento, esta herramienta abierta y de uso libre tiene la capacidad de iluminar la función de cualquier proteína oculta en el "proteoma oscuro" (conjunto de proteínas cuya función aún se desconoce).

Mediante esta novedosa herramienta, llamada FANTASIA (*Functional ANnoTAtion based on embedding space SImilArity*), el equipo del IBE y el CABD ha analizado cerca de 1.000 genomas animales con una precisión cercana al 100%, y ha asignado la función de 24 millones de genes codificantes de proteínas del proteoma oscuro. FANTASIA es capaz de trabajar con *Big Data* para analizar un genoma animal completo en cuestión de horas en un ordenador corriente, o en 30 minutos en un equipo especializado.

Hoy en día se da por hecho que podemos sintetizar insulina para tratar la diabetes, pero esto no sería posible sin comprender la función de esta proteína esencial para la vida. Igual que la insulina, cada proteína cumple una función, y son los genes los encargados de codificarlas, brindando a las células el potencial de expresarlas a través de su maquinaria una y otra vez. El genoma de cualquier organismo alberga la fórmula para sintetizar cualquiera de sus proteínas, es decir, su proteoma. Sin embargo, desconocemos la función de buena parte de los genes que vertebran el árbol de la vida.

En humanos ya se conoce la función de la mayoría de las proteínas —alrededor del 80%-90%—, pero en otros mamíferos esa cifra disminuye, y en invertebrados la función de más de la mitad de las proteínas sigue siendo un misterio. Aunque es posible leer los miles de millones de letras de su secuencia de ADN codificante, la función biológica de muchas de esas proteínas permanece oculta, y con ello se escapan pistas fundamentales sobre la evolución de las especies, su metabolismo o, incluso, su salud. Hasta la fecha, la principal forma de predecir su función era comparando los genes que las codifican con otros similares en su secuencia genética, llamados homólogos, un método limitado que deja fuera buena parte de ese universo aún por explorar.

Descifrando el proteoma oscuro del árbol de la vida de los animales

En la última década, proyectos punteros en el ámbito internacional como el Atlas Europeo de Genomas de Referencia (ERGA por sus siglas en inglés), parte del proyecto BioGenoma de la Tierra (EBP por sus siglas en inglés), han logrado generar secuencias de genomas de referencia de miles de animales para la investigación de la biodiversidad del planeta. Pero acceder a la secuencia que codifica una proteína no significa entender qué hace.

Para desvelar la función de estas proteínas, las metodologías tradicionales (no basadas en IA) comparan los genes que las codifican con secuencias de ADN parecidas, conocidos como genes homólogos. De este modo se *traduce* un proteoma nuevo a partir del parecido con los genes codificantes de otras proteínas ya conocidas. Sin embargo, una

gran mayoría de las proteínas carecen de homólogos de referencia y permanecen ocultas en la *terra ignota* del proteoma oscuro.

"Comprender la función de estos genes gracias a esta nueva herramienta abre una nueva ventana al conocimiento de la biología animal. Nos permitirá entender cómo surgen las innovaciones evolutivas y qué papel desempeñan las proteínas desconocidas en la diversidad y adaptación de las especies", explica **Fernández**, investigadora principal del IBE en el Metazoa Phylogenomics and Genome Evolution Lab y miembro del comité ejecutivo de ERGA.

En esta línea, **Rojas**, que colidera el estudio desde el CABD, subraya que "el uso de modelos de lenguaje basados en inteligencia artificial nos permite ir más allá de la simple comparación por homología. Estos modelos aprenden directamente de las secuencias genéticas y son capaces de inferir la función potencial de genes sin equivalentes conocidos, abriendo nuevas posibilidades para explorar el proteoma oscuro". En este sentido, EBP recomienda el uso de FANTASIA a todos sus colaboradores en la página web del proyecto.

Con los modelos de lenguaje, un tipo específico de aplicación de IA, por primera vez es posible predecir la función de una proteína sin necesidad de comparar la secuencia de sus genes codificantes con la de otros genes conocidos. En lugar de buscar similitudes directas, estos métodos traducen las secuencias de ADN en fragmentos y las analiza *sintácticamente*, como si fueran frases en un idioma. Cada fragmento de la secuencia recibe un valor numérico y, con ellos, el sistema construye su propia *gramática* para anticipar lo que falta, del mismo modo que un procesador de texto completa oraciones.

Este *ChatGPT de las proteínas* aprende de miles de ejemplos ya estudiados, identificando qué hace cada proteína, en qué proceso biológico participa y en qué parte de la célula se encuentra (lo que los científicos llaman términos GO, del inglés *Gene Ontology*). Con esa información, cada proteína se convierte en un vector numérico, una especie de huella digital matemática que resume sus características. Gracias a estos vectores, FANTASIA puede analizar nuevas secuencias de ADN y predecir su función con gran precisión, abriendo la puerta a descubrimientos que antes parecían inalcanzables. Y lo hace con miles de proteínas a la vez.

"FANTASIA es un *software* abierto y fácil de usar para usuarios sin experiencia en programación. Incluye modelos ya entrenados, por lo que se acoge a los principios de sostenibilidad y puede usarse sin necesidad de acceso a superordenadores", comenta Gemma Martínez Redondo, estudiante de doctorado del IBE y primera autora del estudio.

Arrojar luz sobre la "biología oscura": ¿una FANTAS-IA?

Descubrir las funciones que cumplen las proteínas de un organismo es crucial para descifrar la evolución de los genomas y la complejidad de la vida, por lo que este nuevo modelo de lenguaje podría impulsar el conocimiento de la comunidad científica en este campo, pero también en el estudio de la biodiversidad y la salud global.

"FANTASIA es un generador de hipótesis: esta herramienta arroja luz en la oscuridad, puesto que es impensable estudiar todos los genes uno a uno de cada organismo. Ahora, será más fácil dirigir los esfuerzos para investigar en profundidad la función de las proteínas. Esto puede ser muy útil en el ámbito farmacéutico, identificando dianas terapéuticas, por ejemplo," comenta **Fernández**.

El estudio ya ha revelado proteínas ocultas de tardígrados, ctenóforos y micrognatozoos, tres filos invertebrados poco conocidos cuyo proteoma sigue estando oculto en su mayor parte.

"En la biología evolutiva, el cambio, la pérdida o la ganancia de la función de proteínas en los diferentes organismos cuentan la historia de la evolución de su filo o especie. Puede indicarnos cómo se adaptó un organismo a un nuevo medio, de qué se alimentaba o por qué dejó de necesitar ciertas herramientas de su genoma, entre muchas otras", añade **Fernández**.

La herramienta de IA desarrollada se encuentra a disposición de cualquier grupo de investigación en el mundo, con el potencial para iluminar la investigación genómica y proteómica virtualmente en cualquier ámbito de aplicación.

"Sabemos que otros grupos de investigación a nivel internacional ya están utilizando FANTASIA en sus investigaciones, y estamos viendo que no solamente funciona en animales, sino también en plantas, virus, hongos o protistas. El potencial para descubrir nuevos genes que revolucionen la biotecnología, la medicina o la conservación de la biodiversidad no tiene límite", concluye Fernández. "Las posibilidades de los métodos que usamos en esta herramienta son enormes para completar el conocimiento sobre funciones alternativas de proteínas conocidas, es como descifrar su gramática", puntualiza **Rojas**.

Martínez-Redondo, G. I., Perez-Canales, F. M., Carbonetto, B., Fernández, J. M., Barrios-Núñez, I., Vázquez-Valls, M., Cases, I., Rojas, A. M., & Fernández, R. **FANTASIA leverages language models to decode the functional dark proteome across the animal tree of life**. *Communications biology*, *8*(1), 1227. DOI: 10.1038/s42003-025-08651-2

IBE-CSIC Comunicación

comunicacion@csic.es