

# ANOTACIÓN DE FUNCIÓN EN PROTEÍNAS USANDO MODELOS DE LENGUAJE

Ana M Rojas Mendoza\* e Ildefonso Cases

Centro Andaluz de Biología del Desarrollo (CABD), CSIC, Sevilla.  
Grupo de Biología Computacional y Bioinformática

Gemma Martinez-Redondo y Rosa Fernández

Instituto de Biología Evolutiva (IBE), CSIC, Barcelona.  
Metazoa Phylogenomics lab.

El poder transformador de la Inteligencia artificial (IA) ha irrumpido en la biología estructural, mediante desarrollos relacionados con la predicción de la estructura terciaria (ahora de complejos de proteínas y ligandos tras meses de espera), y con el diseño de proteínas a la carta. Dado que la IA ha llegado tarde a la biología, uno esperaría que el impacto sobre todas las disciplinas del dominio biológico sea gradual. Lo que es evidente es que la IA ha llegado para quedarse, y es cuestión de tiempo que llegue a todas las áreas en las que pueda florecer.

La predicción de función es un problema bien conocido, persistente, y que sigue sin resolverse, pese a décadas de desarrollo y esfuerzos los métodos no progresan, representando un reto enorme de la Biología computacional, así que demos un repaso histórico y conceptual a la problemática.

**¿Qué es función?** Es un concepto subjetivo, muy dependiente de contexto, que normalmente extraemos de la literatura científica. Para organizar la información, hemos generado ontologías de genes (Gene Ontology), que incluyen términos GO (por sus siglas en inglés), que no son más que vocabularios controlados relacionando términos asociados a funciones. Estos términos se representan en una estructura de grafo donde los nodos del mismo son los términos, conectados por flechas que indican las relaciones entre los mismos. La raíz del grafo representa los términos más generales (por ejemplo, Proceso Biológico, etc.) y va progresando con términos más específicos si avanzamos de nodo en nodo. La profundidad en el grafo es variable denotando la especificidad del término, y no todas las ramas del grafo tienen la misma profundidad, ni todos los organismos tienen los mismos



grafos. Por ejemplo, los términos relacionados con desarrollo de ala en mosca no se encuentran en el grafo de la levadura.

Lo que anotamos en realidad es el producto codificante de los genes, las proteínas, y esas anotaciones se asocian de nuevo a los genes correspondientes. Por otro lado, "función" también es un concepto bioquímico, y en estructura de proteínas se asocia a sitios de unión o sitios catalíticos. Hay que decir que existe un sesgo en cuanto a los genes que se estudian, puesto que Stoeger *et al.*, en 2018, publicaron un trabajo en el cual demuestran que estudiamos los mismos genes de manera circular, y que incluso en humanos, el foco cae en unos 2.000 genes de los 19.000 posibles. Atribuyeron causas multifactoriales, pero el mensaje es que no estamos progresando. En otras palabras: la información funcional de la que disponemos

es muy parcial y está muy sesgada por intereses específicos.

**¿Cómo se asigna la función a una secuencia nueva?** Pues usamos la similitud de secuencia (concepto matemático, cuantitativo) para identificar homología de secuencia (concepto cualitativo). Esta aproximación está basada en dos premisas, la primera se fundamenta en la evolución molecular y las relaciones evolutivas entre las secuencias de las especies, en particular los ortólogos y los parálogos (ver Ohno, 1970 y Zuckerkandl y Pauling, 1965). Recordamos que un ortólogo es un gen que estaba presente en el ancestro común de dos especies, y que cuando ocurrió el evento de especiación cada especie nueva adquirió una copia del gen, mientras que un parálogo es un gen que se duplica a partir del ortólogo tras el proceso de especiación (Figura 1A).

Y ¿cómo relacionamos esto con la función? Pues a través de la conjetura del ortólogo propuesta por Nehrt *et al.*, en 2011, que sugiere que los ortólogos retienen la función, mientras que los parálogos (duplicaciones de un gen tras especiación) la diversifican, aunque estas nociones las introdujo E. Koonin en los 90. Es decir, asumimos que las secuencias ortólogas "realizan" la misma función. Sin embargo, uno de los errores conceptuales más profundos y persistentes en disciplinas como la biología del desarrollo o la genómica, es utilizar la similitud funcional entre dos genes para definir ortología, como ya discutió G. Theißen en 2002. La analogía funcional no implica ortología, ni la ortología implica necesariamente la analogía funcional.

Desde un punto de vista termodinámico, los estudios seminales de C. Anfinsen en los 60, llevaron a concluir que la estructura más

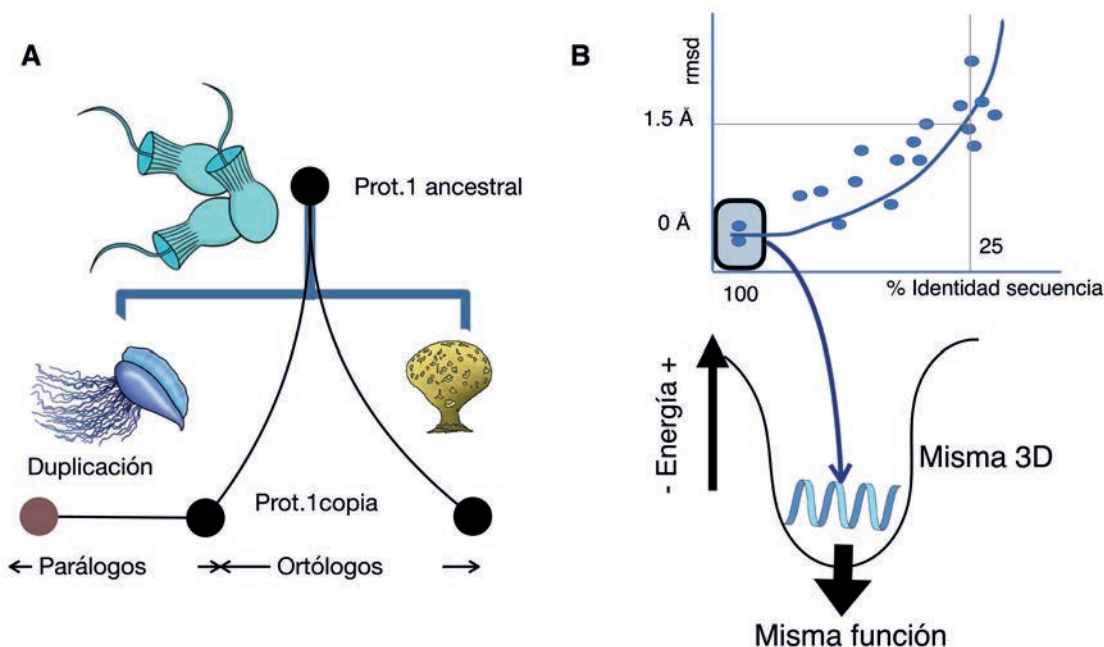


Figura 1

Premisas en las que basamos la asignación de función. A) Relaciones evolutivas a nivel molecular respecto a eventos de especiación. El ancestro común posee un gen (círculo negro) que se transfiere a cada especie tras un evento de especiación. En un caso, se duplica y en otro no. B) El paradigma de secuencia/estructura/función entre la estructura tridimensional y la secuencia, a mayor similitud de secuencia, mayor la de su estructura (el eje Y es *rmsd*, la distancia entre carbonos alfa de la estructura, mientras que el eje X es el % de identidad de la secuencia de aminoácidos). Siguiendo con la hipótesis de Anfinsen, el estado de energía más bajo es el más estable y el que se asocia a la función de la proteína. Hoy sabemos que esto no es así.

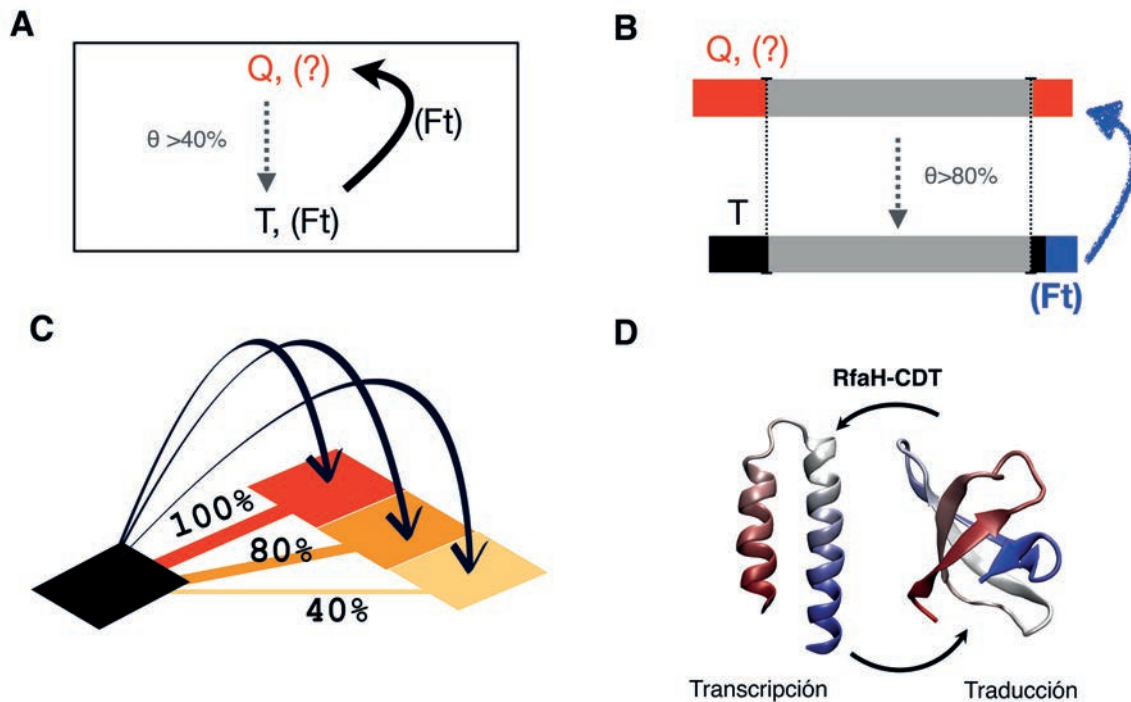


Figura 2

Asignación de función. A) Transferencia de “función” basada en la similitud de la secuencia de aminoácidos. B) Problemas de asignación automática de función por transferencia errónea (función azul). C) Problemas de asignación masiva a miembros de subfamilias de proteínas. D) Caso de metamorfismo en proteínas, que implica cambio radical de estructura cuando la proteína adopta diferentes funciones.

estable asociada a una secuencia se asociaba a una función fisiológica determinada (Figura 1B). Estas dos premisas han originado el siguiente paradigma: “una secuencia, una función, una estructura”, que es central a todos los métodos computacionales de predicción de función y predicción de estructura. Entonces, si dos secuencias de proteínas muestran una identidad mayor a un umbral (normalmente un 40%), que puede obtenerse usando métodos que estiman cómo de idénticas son dos secuencias en su composición (p. ej. BLAST), se transfiere toda la anotación simultáneamente de una a otra (Figura 2A). Esta transferencia es ruidosa, puesto que a veces se transfieren funciones que no están en la secuencia a identificar porque la región responsable de la función (imaginemos un dominio kinasas) no está presente en la secuencia que

queremos anotar (ver en Figura 2B, región azul). También puede ocurrir que se transfiera la misma función a miembros de familias de proteínas, que son aquellas que se dividen en subfamilias a distintos niveles de identidad de secuencia, pero por encima del umbral, que realizan funciones diferentes (Figura 2C). Además, hay numerosas excepciones que invalidan estas premisas, por ejemplo, hay ortólogos que diversifican función, mientras que hay parálogos que adoptan la función original del ortólogo. En otros casos, las proteínas tienen más de una función o la cambian como consecuencia de señales biológicas. Un caso muy llamativo de esto son las histonas, proteínas clásicas en el mantenimiento estructural de la cromatina, que se convierten en enzimas (reducen cobre) cuando se asocian en tetrámeros en ciertas condiciones. Y

no menos relevante, hay proteínas que cambian su estructura sustancialmente para cambiar de función, las llamadas proteínas “metamórficas” (Figura 2D).

Por último, incluso asumiendo que el paradigma fuese cierto, hay miles de proteínas que no presentan homología de secuencia con ninguna otra en las bases de datos, especialmente aquellas que provienen de organismos no modelo, o que tienen una composición rara (baja complejidad), lo que hace que frecuentemente queden sin anotar.

Por lo tanto, el problema subyace en las **premisas**, no en los métodos, que son muy potentes, ni en los datos, cuya disponibilidad es enorme. Sin embargo, la mejora en la predicción se ha estancado, según los resultados publicados del último CAFA (*Critical Assessment of Functional Annotation*), la comunidad que

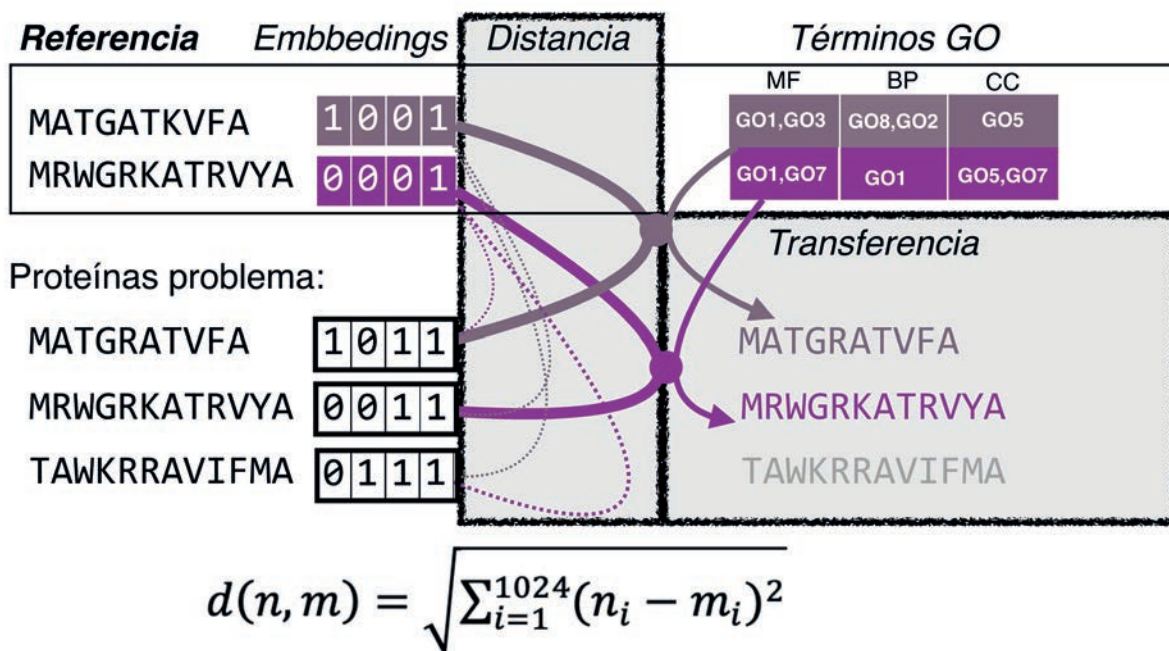


Figura 3

Asignación de función usando modelos de lenguaje. Las secuencias tanto de referencia como problema se codifican en *embeddings* mediante *transformers*, se calcula la distancia en el espacio de *embeddings*, y a distancias más pequeñas (líneas más anchas) se transfieren los términos GO de las secuencias de referencia.

evalúa los métodos de predicción de función\*.

**Entonces, ¿Cómo podemos proceder ignorando el paradigma?**, usando aproximaciones que no requieran la homología de secuencia.

En este contexto, los métodos de IA basados en “*transformers*” emergen como alternativas prometedoras, puesto que no usan información evolutiva.

Los modelos de lenguaje de proteínas (pLM) se han derivado de los modelos de lenguaje natural. En IA, igual que en el lenguaje natural, las frases tienen estructuras, palabras y reglas, y podemos considerar una secuencia de proteínas como una frase y extraer propiedades de esa frase. Hay muchos pLM (introducidos brillantemente en el artículo del Dr. Chacón en este mismo dossier), pero en este artículo nos centraremos en el modelo ProtTrans

publicado por Elnaggar *et al.*, en 2022, que es el que más se ha utilizado para la anotación de función.

Estos modelos que se han adoptado del lenguaje natural, codifican las secuencias de aminoácidos en objetos matemáticos denominados *embeddings*, que son básicamente vectores. Littman *et al.*, en 2021, fueron pioneros en implementar este método para transferir anotaciones, reemplazando la homología de secuencia por la distancia en el espacio de vectores (Figura 3), y usando el set de datos de función de CAFA3, que es un conjunto de datos curados de función, para validar su método. Sin embargo, siendo útiles, estos conjuntos tienen un carácter universal, que no permite el abordaje a nivel de organismo.

Nosotros decidimos abordar estas limitaciones, y para ello en

un esfuerzo colaborativo nos planteamos resolver tres cuestiones fundamentales: ¿Cómo funcionan estos métodos a nivel de organismo? ¿Cómo de bien recuperan la función obtenida en experimentos de expresión génica? Y si los usamos a gran escala en organismos no modelo poco convencionales, ¿Podemos anotarlos con confianza y obtener información funcional relevante?

Para responder a la primera pregunta usamos organismos modelo, que sirven de buena referencia porque tienen anotaciones manuales muy trabajadas y revisadas por expertos (por ejemplo, ratón, mosca, gusano y levadura). Lo primero que nos sorprendió fue comprobar que el estado de anotación de estos organismos, no es completo. Por ejemplo, un 30% de los genes del gusano *C. elegans* carece de anotaciones, al igual que la mosca *D. melanogaster*, con un

17% aproximado de sus genes sin información funcional.

La mayoría de estos genes, o bien no presentan homología identificable con otras secuencias en las bases de datos, o bien estas secuencias presentan características particulares como complejidad, desorden, etc. La idea era reemplazar las anotaciones de referencia, que están bien definidas por las anotaciones predichas por los pLM, y compararlas.

Pero, ¿Cómo evaluamos si las predicciones son buenas? Usamos varias aproximaciones. La primera es valorar su **"precisión"**, es decir, cómo es el acierto de términos. Comparando cuántas veces un método de predicción acierta el término anotado, nos da una idea de la precisión. Otro aspecto a considerar **es el "alcance"** del método. Imaginemos que nuestro método fuese un arma automática, lo relevante es saber cuántas veces tendría que "disparar" para acertar los términos (los GO anotados de referencia). Si dispara mucho, por puro azar acertará alguna vez, pero fallará muchas veces si no es un buen método. Si dispara muy poco y acierta todas las veces, será un método muy preciso, pero con alcance limitado pues no acierta todo lo que podría acertar. Un método perfecto sería aquel que disparando todas las veces que debería disparar (el número real de anotaciones), acertara todas las veces en la diana.

Especialmente relevante, **"cómo de informativo"** es el acierto, en función de dónde esté el término en el grafo de GO. Si el método sólo acierta cerca de la raíz que es el término más general (p. ej. *"Molecular Function"*), entonces la información que proporciona no es detallada y no ayuda a discriminar funciones. Y, por último, evaluando **"cómo de similar"** es la información que produce podemos establecer si

el método está acertando como debería. Para ello, calculamos la similitud semántica de los términos predichos respecto a los anotados por expertos, que es una manera de medir si los grupos de términos asociados a cada proteína son similares, y en su caso, la similitud ha de ser alta para que el método sea fiable.

Estimando todos estos aspectos, concluimos que el modelo ProtTrans anota prácticamente todos los genes y produce más anotaciones por proteína de manera específica, precisa y muy similar.

Es decir, el método recapitula la información obtenida por métodos tradicionales y además anota todo el genoma.

Para responder a la segunda gran pregunta, extrajimos experimentos de transcriptómica publicados para todos los organismos modelo, de los que obtuvimos los genes diferencialmente expresados en distintas condiciones. Calculamos experimentos de enriquecimiento de funciones en esos experimentos de RNAseq para extraer términos enriquecidos. Seguidamente reemplazamos las anotaciones de referencia por las predicciones realizadas por ProtTrans, y repetimos el proceso anteriormente descrito. Con las anotaciones y predicciones, calculamos la similitud semántica entre términos enriquecidos a partir de anotaciones reales *versus* anotaciones predichas obteniendo valores altos, indicativos de una gran similitud.

En otras palabras, las anotaciones predichas capturan la misma información funcional, lo que implica que estos métodos sirven para abordar experimentos asociados a identificación de funciones en organismos no modelo. Todos estos resultados se han publicado en Barrios-Núñez *et al.*, en 2024.

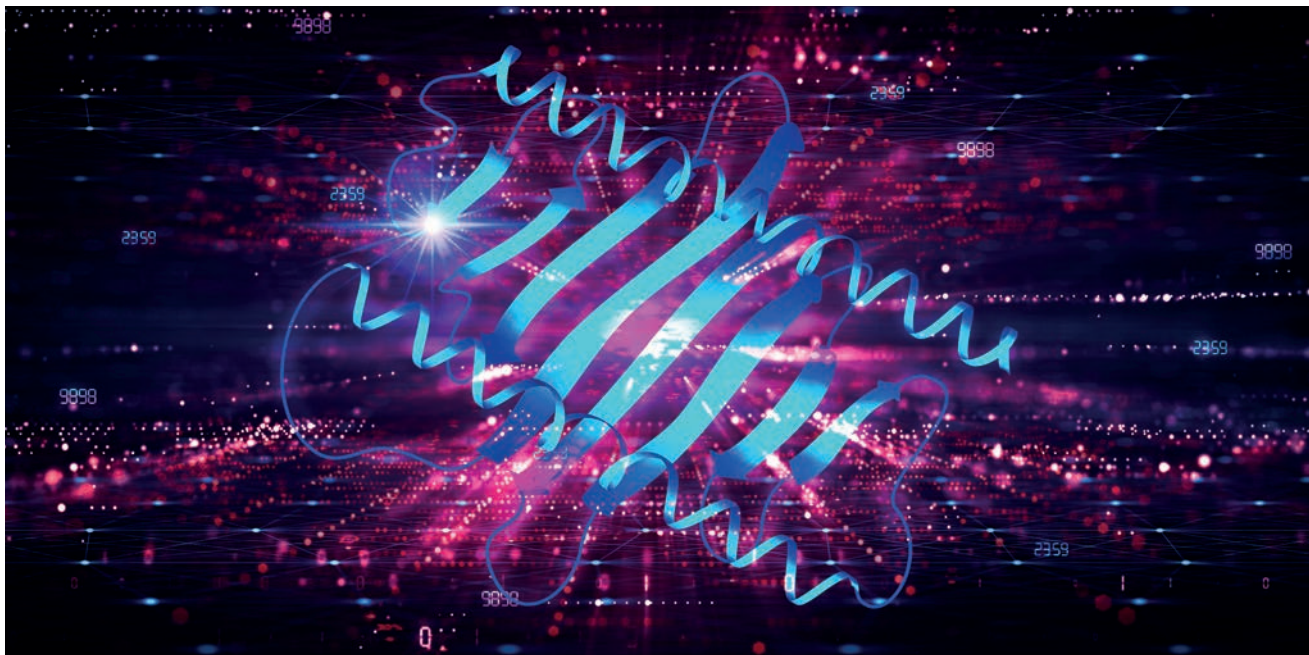
Visto el rendimiento y la fiabilidad de este método, **decidimos**

**usarlo a gran escala**, respondiendo a nuestra última pregunta, y para ello se anotaron unos 24 millones de genes de 1.000 especies de animales basales, invertebrados, muy relevantes para estudios de biodiversidad en procesos de terestrialización. Los métodos tradicionales de referencia anotaban únicamente alrededor de un 50% de los proteomas analizados, dejando en torno a un 50% de media sin anotar, lo que denominamos el "proteoma oscuro".

Los modelos de lenguaje permitieron anotar prácticamente todos los genes de todos los genomas. Cuando analizamos las funciones del proteoma oscuro, observamos enriquecimientos de funciones muy coherentes con la biología de esos organismos. Por ejemplo, en tardígrados, que son unos organismos prácticamente indestructibles, se identificaron genes previamente sin anotar, con funciones asociadas a su indestructibilidad, como términos asociados a resistencia a UV, radiación, calor, etc. Los resultados están accesibles en Martínez-Redondo *et al.*, en un *preprint* 2024. Para facilitar a la comunidad el acceso a una herramienta que permita anotar proteomas de manera masiva, hemos desarrollado FANTASIA.

En el último congreso de ISMB2024, se presentó en el panel *Function-COSI (Community of Special Interest)* el resultado preliminar de la competición en predicción de función CAFA5, donde nueve de los diez mejores métodos usaron modelos de lenguaje.

En conclusión, los modelos de lenguaje de proteínas han llegado para quedarse. Son herramientas muy útiles y alternativas a métodos tradicionales por su capacidad de anotar y por su independencia a criterios evolutivos. Son muy rápidos y no requieren de grandes necesidades computacionales.



## Para leer más

Barrios-Núñez I, *et al.* "Decoding functional proteome information in model organisms using protein language models". *NAR Genomics and Bioinformatics* 6 (2024) lqae078. <https://doi.org/10.1093/nargab/lqae078>

\*CAFA (*Critical Assessment of Functional Annotation*) es un consorcio de científicos expertos en predicción de función que monitorizan el rendimiento de los métodos. Los resultados de estas actividades suelen presentarse en la mayor conferencia de Biología Computacional (ICSB), en la sesión de la comunidad de interés COSI-Function.

Elnaggar A, *et al.* "ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning" en *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 7112-7127 (2022).

FANTASIA (<https://github.com/MetazoaPhylogenomicsLab/FANTASIA>).

Littmann M, *et al.* "Embeddings from deep learning transfer GO annotations beyond homology". *Science Reports* 11 (2021) 1160. <https://doi.org/10.1038/s41598-020-80786-0>

Martínez-Redondo IG, *et al.* "Illuminating the functional landscape of the dark proteome across the Animal Tree of Life through natural language processing models" *bioRxiv* (2024). <https://doi.org/10.1101/2024.02.28.582465>

Nehrt NL, *et al.* "Testing the ortholog conjecture with comparative functional genomic data from mammals". *PLoS Computational Biology* 7 (2011) e1002073. <https://doi.org/10.1371/journal.pcbi.1002073>

Stoeger T, *et al.* "Large-scale investigation of the reasons why potentially important genes are ignored". *Plos Biology* 16 (2018) e2006643. <https://doi.org/10.1371/journal.pbio.2006643>

Theißen G. "Orthology: Secret life of genes". *Nature* 415 (2002) 741. <https://doi.org/10.1038/415741a>