Proteómica y Bioinformática Estructural

Anexo al Manual

Título: *Máster Universitario en Bioinformática*
Materia: *Bioinformática estructural*
Créditos: *6 ECTS*
Código: *07MBIF*
***Edición**: 04_2024*
***Creado por**: Magdalena Nikolaeva Koleva*

4.3.3. Métodos basados en inteligencia artificial

En los últimos años la sociedad está experimentando el auge de diversos métodos generativos de información basados en el desarrollo de algoritmos de inteligencia artificial. El ámbito de la proteómica no ha sido una excepción debido a la necesidad de disponer de estructuras proteicas tridimensionales para abordar una de las aplicaciones más importantes en este ámbito, la búsqueda y desarrollo de compuestos activos para el sector farmacéutico, entre otros. De esta forma surgen **AlphaFold** y **AlphaFold2**, modelos de inteligencia artificial generativa basados en aprendizaje automático por redes neuronales a partir de estructuras proteicas conocidas.

Antes de profundizar en el funcionamiento de AlphaFold deberíamos definir los términos de ingeniería informática en los que se basa: Inteligencia artificial (*Artificial Intelligence - AI*), aprendizaje automático (*Machine Learning - ML*) y aprendizaje profundo (*Deep Learning - DL*).

a. **Inteligencia artificial**: es el estudio que tiene como objetivo crear máquinas o programas que imiten la capacidad cognitiva humana para realizar diversas tareas, como el aprendizaje, la resolución de problemas, la comprensión del lenguaje y la toma de decisiones.
b. **Aprendizaje automático**: es un subapartado de la inteligencia artificial centrado en el desarrollo de algoritmos que pueden aprender a partir de conjuntos de datos para realizar predicciones o tomar decisiones en base a ese aprendizaje. En otras palabras, se trataría de enseñar a una máquina a aprender desde la experiencia, como los humanos. Existen distintas formas de aprendizaje automático; **supervisado**, donde los datos empleados para el aprendizaje ya están etiquetados, **no supervisado**, donde el entrenamiento se realiza con datos sin etiquetar y por tanto el algoritmo debe sacar sus propias conclusiones y formas de clasificar dichos datos, y por **recompensa**, donde el algoritmo aprende en base a recompensas o castigos por el entrenador.
c. **Aprendizaje profundo**: es un subapartado del aprendizaje automático centrado en el desarrollo de redes neuronales artificiales que aprenden a partir de grandes conjuntos de datos. En realidad, se inspira en la estructura y función del cerebro humano, con múltiples capas de nodos interconectados que procesan la información.

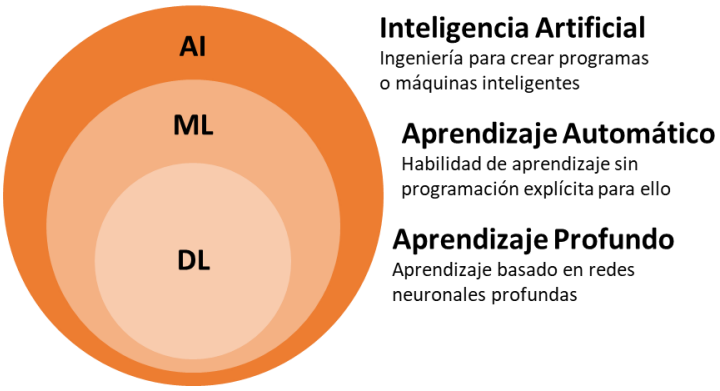

Figura 1. Campo de inteligencia artificial y los distintos subapartados que abarca.

Ahora que conocemos dónde se posiciona AlphaFold, veamos cómo funciona y en qué se basan sus predicciones estructurales a partir de secuencias aminoacídicas. El método que emplea se puede dividir en 3 etapas principales: (i) preprocesamiento y búsqueda en bases de datos, (ii) aprendizaje profundo (*Evoformer*) y (iii) modulo estructural y postprocesamiento.

a. El **preprocesamiento** se divide en dos etapas que permiten generar distinta información a partir de la secuencia que se desea modelar. La primera se basa en obtener el alineamiento de secuencia múltiple (*multiple sequence alignment - MSA*) empleando la secuencia de aminoácidos de entrada y comparándola con secuencias de distintas especies de proteínas similares (homólogas), evolutivamente cercanas y por tanto con secuencias parecidas. De esta forma si hay una mutación en cualquier residuo debe haber una mutación equivalente en otro residuo que conserve la interacción con el primero dentro de la propia proteína para mantener su función. Por tanto, podemos decir, que estos aminoácidos coevolucionan. Ello permite inferir qué interacciones se dan dentro de la propia proteína y como resultado el plegamiento de la misma.

   Estos resultados a su vez permiten la generación de histogramas bidimensionales que representan de forma bidimensional las distancias entre cada aminoácido dentro de la proteína para obtener una idea de su representación en 3D, dónde están más cerca y por tanto puedan interaccionar. A continuación, el algoritmo buscaría similitudes entre los histogramas de estructuras proteicas conocidas y publicadas en la base de datos *Protein Data Bank* y la secuencia de entrada que se quiere modelar.

b. La segunda parte del modelo consiste en la comunicación entre los dos parámetros anteriores mediante el empleo de una red neuronal artificial de aprendizaje profundo (***Evoformer***) que modifica los pesos de las funciones que conectan cada nodo para modificar las salidas o resultados obtenidos tras aplicar el algoritmo conforme va aprendiendo. Se llevan a cabo 48 bloques de optimización antes de producir un resultado.

c. Finalmente, encontramos el **módulo estructural** que involucra otra red neuronal que realiza rotaciones y traslaciones sobre los aminoácidos de la secuencia para crear un primer modelo estructural de la proteína. Aplica diversos tipos de restricciones de tipo físico y químico dictados por los enlaces, ángulos de enlace y ángulos de torsión refinando el modelo. La secuencia completa desde el *Evoformer* hasta la obtención del modelo estructural se repite tres veces para generar un resultado optimizado.

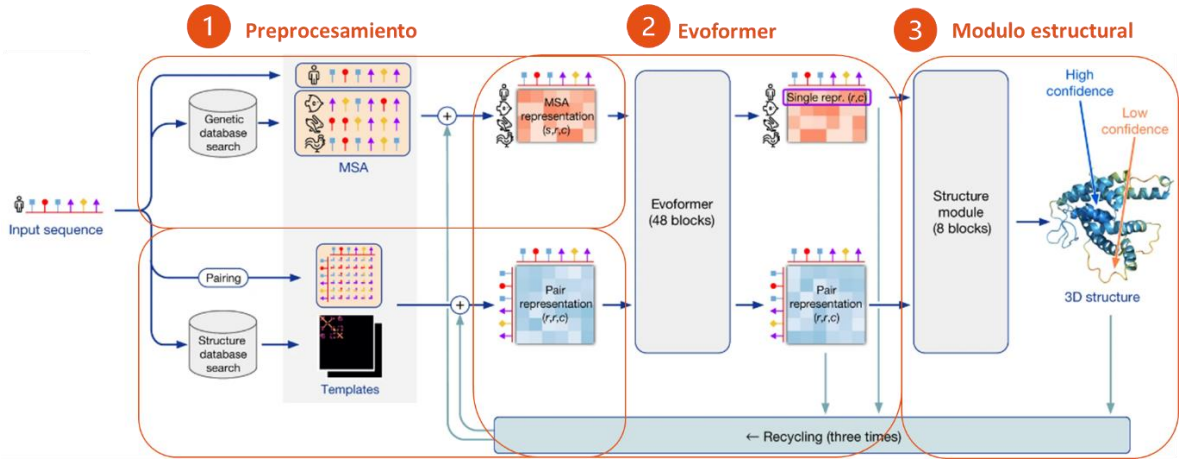

Figura 2. Principales etapas del modelo predictivo de AlphaFold. Adaptado de Jumper J. et al, Nature, 2021.

En resumen, podríamos decir que AlphaFold se basa en la **evolución** ya que aprende de las relaciones entre **proteínas homólogas** para comprender cómo las **mutaciones** en la secuencia afectan a la estructura proteica y que su funcionamiento gira alrededor del componente principal que es su **red neuronal artificial de aprendizaje profundo** denominada ***Evoformer***.

7.4. Plegamiento inverso

El **plegamiento inverso** constituye una herramienta potente e innovadora que está transformando el diseño *de novo* de proteínas. Este método ya ha sido empleado para el diseño de agentes terapéuticos, biosensores y enzimas industriales.

A estas alturas de la asignatura ya conocemos qué es el plegamiento proteico, cómo tiene lugar y algunas de las formas de predicción desarrolladas en los últimos años para determinar qué plegamiento 3D tendría una secuencia de aminoácidos, por ejemplo, AlphaFold. Sin embargo, no hemos discutido el proceso contrario, es decir, intentar predecir una secuencia aminoacídica que pudiera plegarse en una estructura 3D determinada, creando una **interacción con una**

diana particular, conteniendo un **sitio catalítico** característico o simplemente presentando una **topología determinada** tras su plegamiento.

Los modelos de plegamiento inverso suelen presentar la ventaja de ser muy rápidos pudiendo predecir en pocos minutos cientos de secuencias que podrían plegarse en la proteína de interés. Además, la identidad entre secuencias suele ser entre 40 y 75% lo cual abre el espacio secuencial explorado, a diferencia de los métodos tradicionales que se basan en pocas mutaciones puntuales. Así mismo, estos métodos permiten fijar regiones de interés, para su conservación, sobre todo en el caso del diseño de péptidos terapéuticos o pequeñas proteínas. Algunos ejemplos de métodos de plegamiento inverso son **RFdifussion** y **ProteinMPNN**:

El método **RFdifussion** se entrena a partir de **estructuras proteicas enmascaradas** cuya secuencia es desconocida para el modelo. Para enmascarar las estructuras se eliminan todas las cadenas laterales de sus residuos, manteniéndose solamente la información estructural de los carbonos α y β, así como de los nitrógenos. Adicionalmente, se introduce **ruido y variabilidad al azar** en las coordenadas de los átomos disponibles para conseguir mayor robustez y evitar el "*overfitting*". El funcionamiento de RFdifussion se divide en 4 pasos: i) definición de la estructura objetivo donde se define la estructura 3D que queremos que tenga la proteína, ya sea empleando una estructura conocida como referencia o diseñando una nueva estructura desde cero; ii) generación de una nube de proteínas donde se utiliza un **modelo de difusión**, un tipo de **modelo de aprendizaje automático** que se utiliza para **generar imágenes** y otros tipos de datos, para generar una gran cantidad de proteínas con diferentes secuencias de aminoácidos; iii) selección de las mejores proteínas mediante el empleo de RoseTTAFold para evaluar la estructura 3D de cada proteína en la nube. Se seleccionan las proteínas que mejor se ajustan a la estructura objetivo; iv) optimización de las proteínas para mejorar su estabilidad y función.

Por otro lado, **ProteinMPNN** basa su predicción en el empleo del **aprendizaje profundo** y las **redes neuronales artificiales**, en este caso **autoregresivas**. En el caso del diseño de proteínas, la autoregresión permite a la red neuronal generar una secuencia de aminoácidos que se ajuste a la estructura 3D deseada. La red "recuerda" los aminoácidos que ya ha seleccionado y elige los siguientes de forma que la secuencia sea compatible con la estructura. Ello proporciona algunas ventajas como la **precisión** ya que se generan secuencias más precisas que en las redes neuronales tradicionales ya que se tiene en cuenta el contexto, y la **eficiencia**, ya que no necesitan procesar toda la entrada de una vez. ProteinMPNN surge para dar una solución a los métodos "clásicos" que predecían la secuencia tratando de minimizar la energía resultante de su plegamiento y que por tanto eran muy lentos y costosos computacionalmente. La red neuronal empleada por ProteinMPNN se caracteriza por presentar 3 capas **codificadoras** y 3 capas **decodificadoras**, donde cada módulo tiene 128 dimensiones

ocultas. En una red neuronal, las capas codificador y decodificador trabajan juntas para realizar una tarea específica. El codificador toma la entrada y la transforma en una representación interna, mientras que el decodificador toma esa representación y la convierte en la salida deseada. Como entrada se usan las características de la estructura del esqueleto peptídico de la proteína como por ejemplo, las distancias entre átomos de carbono α, orientaciones y rotaciones relativas entre carbonos α adyacentes y ángulos diedros del esqueleto peptídico. Como salida se obtiene la secuencia aminoacídica que produciría la estructura proteica de interés.

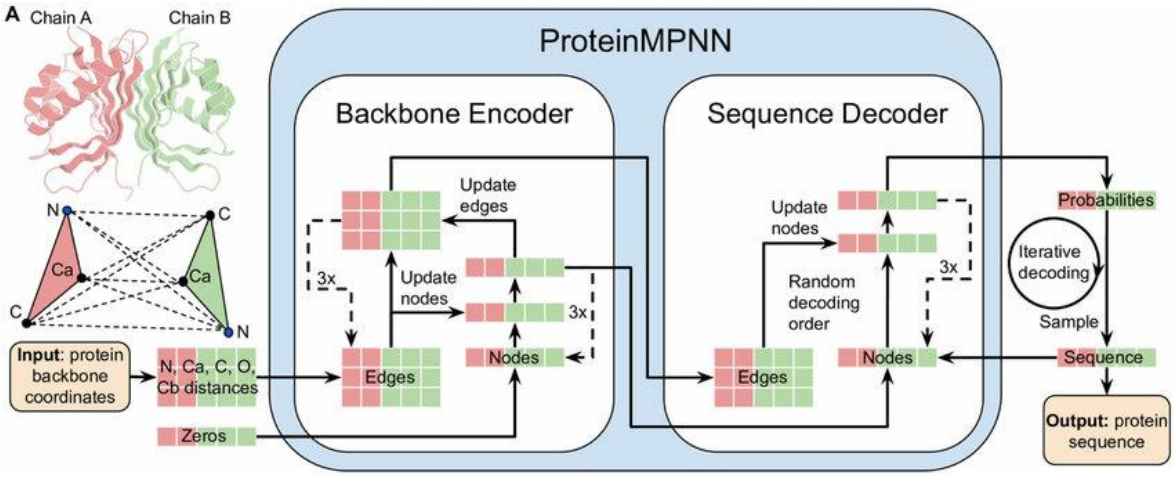

Figura 3. Red neuronal artificial empleada por ProteinMPNN para realizar plegamiento inverso. Consta de las capas codificadoras que reciben la entrada de las características estructurales de la proteína de interés y las capas decodificadoras autoregresivas empleadas para proporcionar una secuencia de aminoácidos como resultado. Adaptado de Dauparas J. et al, Science, 2022.

Bibliografía:

Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

Fuchs F. and Dauparas J. AlphaFold 2 & Equivariance. https://dauparas.github.io/post/af2/

Watson, J.L., Juergens, D., Bennett, N.R. et al. De novo design of protein structure and function with RFdiffusion. Nature 620, 1089–1100 (2023). https://doi.org/10.1038/s41586-023-06415-8

J. Dauparas et al., Robust deep learning–based protein sequence design using ProteinMPNN. Science 378, 49-56 (2022). https://doi.org/10.1126/science.add2187