UNIVERSIDAD NACIONAL DE COLOMBIA

AlphaFold
Introducción al Aprendizaje de Máquina
2021-I

Lizeth Catherine Ortiz Pulido

AlphaFold2 nace como una solución al modelamiento en 3D de las proteínas desarrollado por DeepMind y gracias a la competencia CASP14 del 2020, que consistía en que dada una secuencia que pertenece a una proteína cuya estructura es desconocida, los biólogos computacionales tienen que generar métodos de modelado que puedan predecir esta estructura.

Antes que nada, tenemos que hablar un poco de la importancia del problema en el campo de la biología. Las proteínas son secuencias de aminoácidos y son las moléculas más complejas y sofisticadas que se conocen debido a su plegamiento. Cada tipo de proteína tiene una estructura tridimensional única que viene marcada por su secuencia de aminoácidos. Las proteínas son esenciales para la vida, y el entendimiento de su estructura puede facilitar una comprensión mecanicista de su función. A través de un enorme esfuerzo experimental, se han determinado las estructuras de alrededor de 100.000 proteínas únicas, pero esto representa una pequeña fracción de los miles de millones de secuencias de proteínas conocidas.

Predecir la estructura 3D de una proteína basándose únicamente en su secuencia de aminoácidos, más específicamente el "problema del plegamiento de proteínas", ha sido un importante problema de investigación abierto durante más de 50 años, ya que el método tradicional es muy costoso y demorado. Alphafold2 proporciona el primer método computacional que puede predecir regularmente estructuras de proteínas con precisión atómica, incluso cuando no se conoce una estructura similar.

AlphaFold mejora en gran medida la precisión de la predicción de la estructura mediante la incorporación de nuevas arquitecturas de redes neuronales y procedimientos de entrenamiento basados en las restricciones evolutivas, físicas y geométricas de la estructura de las proteínas, es decir con enfoques bioinformáticos y físicos.

Funcionamiento de AlphaFold2: En primer lugar, el sistema AlphaFold2 utiliza la secuencia de aminoácidos de entrada para construir un alineamiento de secuencia múltiple (MSA), que es una técnica para alinear y comparar muchísimas secuencias de aminoácidos de proteínas de distintos seres vivos que tienen una relación evolutiva y por tanto su secuencia de aminoácidos es parecida. Al hacer esto es posible que podamos detectar correlaciones entre las proteínas y cómo los aminoácidos interaccionan dentro de la proteína. Esa interacción es necesaria para mantener la estructura de la proteína. Como resultado obtenemos una matriz de correlación como esta, es decir una imagen.

En la segunda parte, AlphaFold2 toma la alineación de secuencia múltiple y coordenadas de aminoácidos de un pequeño número de estructuras homólogas disponibles, y las pasa a través de un Transformer, (un modelo de aprendizaje profundo que se usa mucho en modelos de aprendizaje de lenguajes). Gracias a 48 bloques tenemos como resultado un Distograma representando las distancias entre aminoácidos como este, siendo el amarillo muy cercano y azul lejano.

Por último, se toma el distograma o matriz de distancia y con ayuda de un sistema heurístico se produce la predicción final de coordenadas 3D, es decir que para este último paso no se usa aprendizaje

de máquina. A pesar de la larga historia de aplicación de redes neuronales a la predicción de estructuras, solo recientemente han llegado a mejorar la predicción de estructuras. Estos enfoques aprovechan eficazmente la rápida mejora en los sistemas de visión por computadora al tratar el problema de la predicción de la estructura de la proteína como la conversión de una "imagen" de acoplamientos evolutivos en una "imagen" de la matriz de distancia de proteínas.

En la primera versión de AlphaFold presentado en la anterior competencia del 2018 se utilizaban Redes neuronales convolucionales y luego el Algoritmo del descenso del gradiente, algoritmo de optimización aplicado en la estructura de la proteína perfeccionando los ángulos de torción.

Para medir la precisión de los modelos presentados en la competencia se usa el GDT, Global Distance Test, que es el porcentaje de aminoácidos en la posición correcta. Más de 90 GDT se considera comparable a los tradicionales y Alphafold2 obtuvo un puntaje de 92,4 GDT.

Si bien AlphaFold tiene una alta precisión en la gran mayoría de las estructuras de Protein Data Bank depositadas (es una base de datos de la estructura tridimensional de las proteínas y ácidos nucleicos. Estos datos, generalmente obtenidos mediante cristalografía de rayos X o resonancia magnética nuclear, son enviados por biólogos y bioquímicos de todo el mundo), se observa que todavía hay factores que afectan la precisión o limitan la aplicabilidad del modelo. El modelo utiliza múltiples alineaciones de secuencia y la precisión disminuye sustancialmente cuando la profundidad media de alineación es inferior a aproximadamente 30 secuencias. En AlphaFold, se asumen que las proteínas están solas, pero casi nunca están solas, están pegadas y entender cómo se juntan es muy importante, pero hay pocos datos.

Con las bases de datos de las secuencias y esta herramienta, se espera acelerar el avance de la bioinformática estructural que puede seguir el ritmo de la revolución genómica. Se espera que AlphaFold, y los enfoques computacionales que aplican sus técnicas para otros problemas biofísicos, se conviertan en herramientas esenciales de la biología moderna.

Una forma de saber si un desarrollo científico tiene éxito consiste en poderlo usar todo o en parte en la solución de situaciones cotidianas o que tenga impacto en la sociedad, el uso de este sistema se usó para ayudar a que se crearan las vacunas contra el SARS-CoV-2, causante del Covid-19.

Así mismo se podrían crear nuevos fármacos que dependa de proteínas dado que con los métodos tradicionales no se puede conocer la estructura, o que se demora mucho y cuesta mucho. Se está avanzando, aún falta mucho, pero se espera que en unos años estos avances científicos ayuden a la medicina, a la biología y a todas sus aplicaciones.