



UNIVERSIDAD PERUANA
CAYETANO HEREDIA
FACULTAD DE CIENCIAS Y FILOSOFÍA

PRÁCTICA 14 – ALPHAFOLD Y ESMFOLD

Introducción:

Nos encontramos en una era de revolución en la biología estructural, impulsada por el advenimiento de tecnologías innovadoras como AlphaFold, una herramienta pionera en la predicción estructural de proteínas. AlphaFold y otras herramientas de inteligencia artificial han transformado radicalmente nuestra capacidad para determinar la estructura tridimensional de las proteínas, reduciendo drásticamente el tiempo requerido para obtener esta información vital, que antes podía tomar años o incluso la duración de un doctorado.

La capacidad de predecir con precisión la estructura de las proteínas es esencial, pero igualmente crítico es tener métricas que nos indiquen el grado de confianza en estos resultados. Esto permite una evaluación más rápida y precisa de la utilidad potencial de un modelo. Las aplicaciones de AlphaFold son amplias y variadas, desde la generación de modelos para interacciones proteína-proteína, la predicción de regiones intrínsecamente desordenadas. Cabe señalar que el éxito de AlphaFold no habría sido posible sin la vasta recopilación de estructuras cristalográficas acumulada a lo largo de muchos años.

En el avance reciente de los modelos de lenguaje a gran escala, hemos visto la emergencia de capacidades inéditas que trascienden la simple coincidencia de patrones, incluyendo razonamiento de alto nivel y generación de imágenes y textos de aspecto realista. Si bien los modelos de lenguaje basados en secuencias proteicas han sido explorados en una escala más pequeña, existe un vacío de conocimiento en cuanto a lo que aprenden sobre la biología a medida que se aumentan de tamaño. ESMfold ha escalado esta frontera, entrenando modelos de hasta 15 mil millones de parámetros. Se descubrió que al incrementar la escala de los modelos, estos adquieren la capacidad de predecir la estructura tridimensional de una proteína con la precisión de los átomos individuales. En este sentido, ESMFold es una herramienta para la predicción precisa de estructuras atómicas a partir de la secuencia individual de una proteína. ESMFold se equipara en exactitud a herramientas existentes como AlphaFold2 y RoseTTAFold, especialmente para secuencias de baja perplejidad que son adecuadamente procesadas por el modelo de lenguaje.

Objetivo general:

- Entender cómo ingresar adecuadamente la secuencia de aminoácidos como input en AlphaFold.
- Familiarizarse con las opciones avanzadas disponibles en AlphaFold y cómo pueden afectar las predicciones.

- Comprender cómo interpretar los resultados producidos por AlphaFold, utilizando las métricas de Local Distance Difference Test (LDDT) y Predicted Aligned Error (PAE).
- Comprender cómo interpretar los resultados producidos por ESMfold, utilizando las métricas de Local Distance Difference Test (LDDT) y Predicted Aligned Error (PAE).
- Saber cómo visualizar los resultados utilizando el software ChimeraX.

Objetivo específicos:

- Ingresar una secuencia de aminoácidos o varias secuencias de aminoácidos para la predicción de un multímero en AlphaFold.
- Utilizar los diferentes modos de alineamiento de secuencias múltiples (MSA) disponibles en AlphaFold.
- Seleccionar la opción de modelo adecuada para sus necesidades.
- Comprender cómo ajustar el número de reciclajes en AlphaFold para mejorar la calidad de las predicciones.
- Modelar una estructura monomérica con ESMfold
- Predecir la estructura multimérica con ESMfold y evaluar los resultados con LDDT Y PAE.
- Interpretar y utilizar las métricas de LDDT y PAE para evaluar la calidad de las predicciones de estructuras de proteínas.
- Utilizar el software ChimeraX para visualizar de manera eficiente los modelos de proteínas producidos por AlphaFold.

Programa a utilizar :

- Alphafold/ColabFold
- ESMFold

A. Alphafold/ColabFold

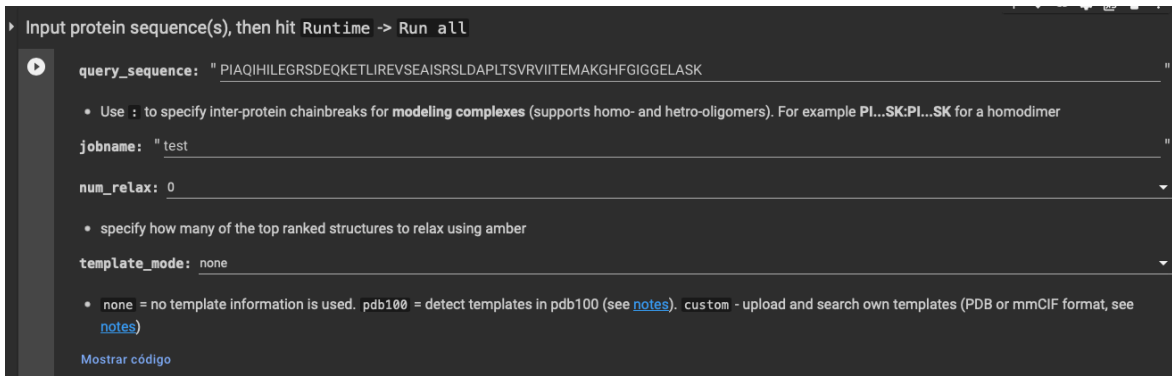
0. Uso de Google Colab para ColabFold:

Para acceder a el Google Colab:

- Solo necesitas abrir el link proporcionado.
 - Para correr:
<https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb>
- Una vez en él, asegúrate de estar conectado a una GPU (puedes verificarlo en el menú de configuración).
- Para ejecutar cada celda de código, simplemente debes hacer clic en el botón de "play" que encontrarás a la izquierda de cada celda.

1. Definición de la Secuencia Input de la Proteína

1. Query Sequence: Ingrese la secuencia de aminoácidos como input.
2. Predicción de Multímero: Si su objetivo es correr una predicción para un multímero, es decir, más de una secuencia de aminoácidos para evaluar la predicción de su estructura compleja, debe separar cada secuencia adicional con dos puntos. Ejemplo: Input_Secuencia1:Input_Secuencia2
3. JobName: Asigne un nombre representativo a su trabajo. Esto facilitará la identificación de qué proteína o conjunto de proteínas se procesaron, ya sea un monómero o un multímero.
4. Relax Number: Especifique si se utilizará la relajación por Amber. Este es un proceso de minimización de energía rápido para resolver conflictos o violaciones estructurales en la estructura de la proteína predicha. La relajación utiliza el campo de fuerza Amber, afectando principalmente a los residuos y mínimamente a la cadena backbone de la proteína.
5. Template Mode: AlphaFold admite inputs que incluyen un template, una estructura modelo de proteína. Para esto, tiene dos opciones: puede elegir PDB100, una base de datos de PDBs agrupada para evitar redundancia de secuencias idénticas, o puede utilizar su propia base de datos personalizada.



Input protein sequence(s), then hit Runtime -> Run all

query_sequence: "PIAQIHILEGRSDEQKETLIREVSEAIRSLDAPLTSVRVIITEMAKGHFGIGGELASK"

• Use : to specify inter-protein chainbreaks for **modeling complexes** (supports homo- and hetro-oligomers). For example PL...SK:PL...SK for a homodimer

jobname: "test"

num_relax: 0

• specify how many of the top ranked structures to relax using amber

template_mode: none

• none = no template information is used. pdb100 = detect templates in pdb100 (see [notes](#)). custom - upload and search own templates (PDB or mmCIF format, see [notes](#))

Mostrar código

2. Opciones de MSA (Multiple Sequence Alignment)

Los alineamientos de secuencias múltiples, o MSA, son cruciales en el proceso de modelado de AlphaFold, ya que se basa en la señal de coevolución encontrada en estos alineamientos para realizar predicciones de alta calidad. La profundidad en la construcción del MSA es un aspecto esencial para obtener buenos resultados. En este contexto, la profundidad se refiere a la cantidad de secuencias homólogas que se pueden encontrar en la base de datos para construir el MSA.

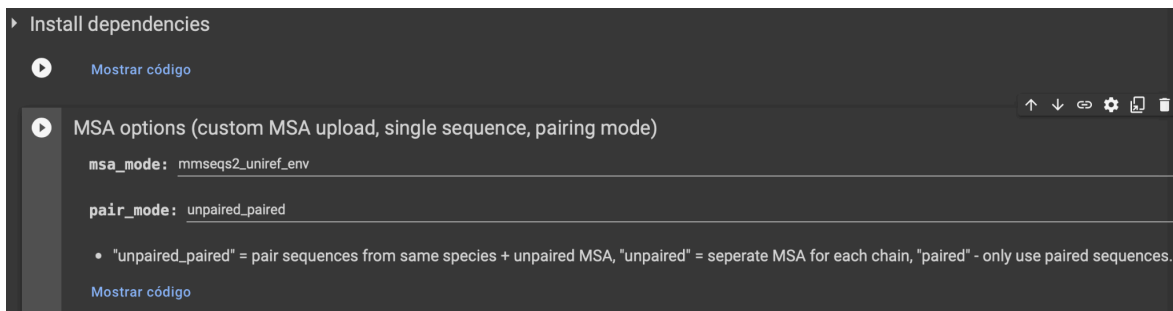
Por ejemplo, ColabFold, la plataforma que estamos utilizando, emplea un método de construcción de MSA rápido que ha demostrado un rendimiento similar al AlphaFold original, que usa el enfoque más lento de Jackhmmer. Sin embargo, el método de ColabFold puede producir un MSA de calidad ligeramente inferior ya que utiliza los algoritmos MMSEQS2 y HHSearch. Una profundidad mínima de 30 en el MSA es

recomendada para AlphaFold. Predicciones basadas en MSAs menos profundos pueden ser menos confiables.

1. Modo MSA: La primera opción que se presenta en el Jupyter Notebook es la del modo MSA. La elección recomendada aquí es MMSEQS2-UNIREF-ENV. Esto indica que se usará el algoritmo MMSEQS2 en las bases de datos UNIREF100, que agrupa todas las secuencias de aminoácidos de Uniprot de manera que no hay secuencias idénticas, reduciendo así la carga computacional innecesaria. 'ENV' indica que se incluirán también todas las bases de datos metagenómicas, que son las más grandes y permiten construir un MSA más profundo, ya que pueden proporcionar más secuencias homólogas. Si se prefiere menos volumen de datos, se puede elegir la segunda opción, MMSEQS2-UNIREF.

2. Modo Single Sequence: Si se opta por no usar MSAs y se prefiere utilizar AlphaFold como un método completamente ab initio (es decir, sin MSAs), se puede seleccionar el modo Single Sequence. Este modo proporciona por defecto un MSA de una sola secuencia y no utilizará ninguna señal de coevolución.

3. Modo Custom: Este modo permite al usuario proporcionar un MSA personalizado si tiene la capacidad de construir un MSA de mejor calidad utilizando algoritmos propios.

A screenshot of a Jupyter Notebook interface. At the top, there is a tab labeled 'Install dependencies' with a play button icon and a link 'Mostrar código'. Below this, there is a section titled 'MSA options (custom MSA upload, single sequence, pairing mode)' with a play button icon and a toolbar with icons for up, down, search, settings, and a trash can. The code in the cell is as follows:

```
msa_mode: mmseqs2_uniref_env  
  
pair_mode: unpaired_paired
```

Below the code, there is a bullet point explaining the pair_mode options: "unpaired_paired" = pair sequences from same species + unpaired MSA, "unpaired" = separate MSA for each chain, "paired" - only use paired sequences. At the bottom of the cell, there is a link 'Mostrar código'.

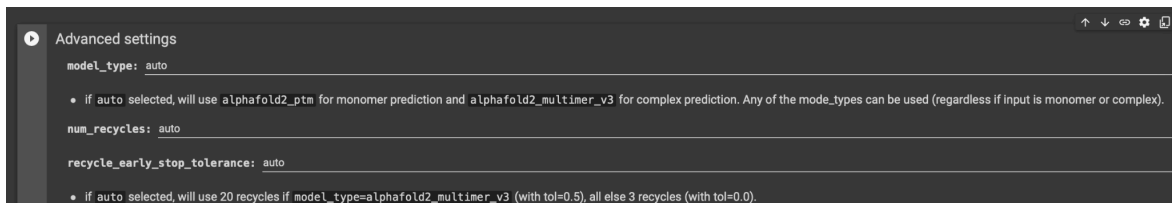
3. Opciones avanzadas

Sección 3: Opciones avanzadas

1. Elección del modelo: Aquí podemos encontrar varias opciones de modelos. Uno es el "monómero", que se utilizará cuando se quiera modelar una única proteína. Para modelar complejos proteicos, AlphaFold ofrece versiones multímero V1, V2 y V3. Es recomendable utilizar la última versión, la V3, ya que se han corregido varios errores, incluyendo la predicción incorrecta de estructuras colapsadas entre otras. Vale la pena recordar que el problema de la predicción de interacciones entre proteínas aún no está completamente resuelto, a diferencia de la predicción de estructuras de monómeros que está cerca. También es importante notar que AlphaFold Multimer es diferente a AlphaFold Monomer. El modelo multímero ha sido entrenado desde cero para la predicción de interacciones proteína-proteína y tiene pesos muy diferentes a los del modelo

clásico de AlphaFold 2-PTM. Si se selecciona "auto", el modelo adecuado se seleccionará automáticamente.

2. Número de reciclajes: El reciclaje se refiere al número de iteraciones que se permite al modelo para refinar su estructura final. En concreto, una vez que se realiza una predicción y se obtiene un PDB con las coordenadas de los átomos de la estructura final predicha, esta estructura se reutiliza como input en el modelo de AlphaFold, reiniciando el ciclo a través del EvoFormer y el Modo Estructural. A veces, la calidad de la predicción mejora significativamente al aumentar el número de reciclajes. El valor predeterminado son tres reciclajes, pero uno puede optar por incrementar este número a 12, 24, 48, o incluso establecerlo en "auto", lo que permitirá al modelo determinar el número de reciclajes en función de cuánto está mejorando la predicción. Esto implica una especie de mecanismo de tolerancia que detiene el reciclaje cuando las mejoras son marginales. Cabe recordar que aumentar el número de reciclajes aumentará significativamente el tiempo de predicción.



4. Interpretación de Resultados

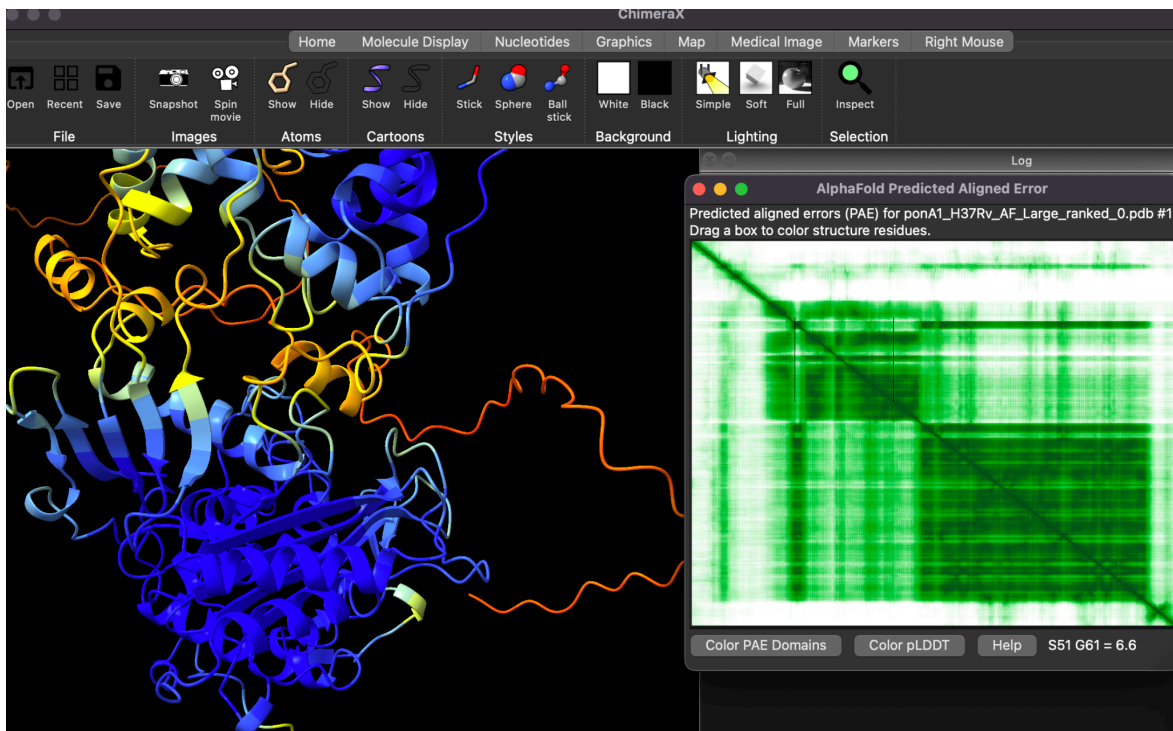
1. Local Distance Difference Test (LDDT):

La métrica LDDT es una herramienta útil para evaluar la precisión de las predicciones de estructuras de proteínas a nivel local. Este enfoque mide la diferencia entre las distancias de todos los pares de átomos en la estructura de la proteína predicha en comparación con la estructura real. Es decir, calcula la discrepancia entre las distancias atómicas correspondientes en el modelo y en la estructura de referencia. Esta métrica proporciona un espectro de puntuaciones de precisión a nivel de residuo que se pueden utilizar para generar mapas de confianza para las predicciones de estructuras de proteínas. Una característica esencial de LDDT es que no requiere un alineamiento previo de las estructuras y puede manejar desplazamientos de dominio, lo que la hace útil para la evaluación de las predicciones de AlphaFold.

2. Predicted Aligned Error (PAE):

La métrica PAE es otra herramienta valiosa para evaluar la calidad de una predicción de estructura de proteína. Esta métrica ofrece una estimación del error en la posición de cada residuo en la estructura predicha, en términos de su orientación relativa con respecto a los otros residuos. En esencia, mide cuán seguros estamos de la posición de un residuo en el espacio tridimensional. Un valor de PAE más bajo indica un mayor grado de confianza en la posición predicha de un residuo, lo que significa que se espera que esté más cerca de la orientación correcta. Esta métrica es especialmente útil para evaluar si los

dominios en la estructura final están correctamente orientados. Visualmente, los residuos con PAE más bajos se representan con colores más oscuros (verde o azul), mientras que los residuos con mayor error se muestran en rojo. Esto proporciona una representación gráfica intuitiva de las áreas de la estructura donde podemos tener mayor o menor confianza en la precisión de la predicción.



5. Visualización de resultados

Sección 5: Visualización de Resultados

1. Descarga el software ChimeraX visitando su página web, seleccionando "Other Releases" y eligiendo la versión que corresponda a tu sistema operativo (MacOS, Windows o Ubuntu/Linux). Ejecuta el instalador y abre el software.

<https://www.cgl.ucsf.edu/chimerax/download.html>

2. Carga tu modelo de proteína en ChimeraX. Puedes hacer esto haciendo doble clic en el archivo PDB que deseas visualizar. Si el software no lo reconoce automáticamente, puedes hacer clic en "Open" y cargar el archivo PDB manualmente.

<https://drive.google.com/drive/folders/1CHbtU0V-MvF5uk63pegg62fUr36X6SrK?usp=sharing>

3. Interactúa con el modelo en 3D haciendo clic y moviéndolo. Si haces clic en un residuo, el software te proporcionará información sobre ese residuo, como el tipo de aminoácido y su posición en la secuencia de la proteína.

4. Visualiza las métricas de tu modelo usando ChimeraX. Ve al menú de "Herramientas" y selecciona "Predicción Estructural" o "Structure Prediction". Haz clic en

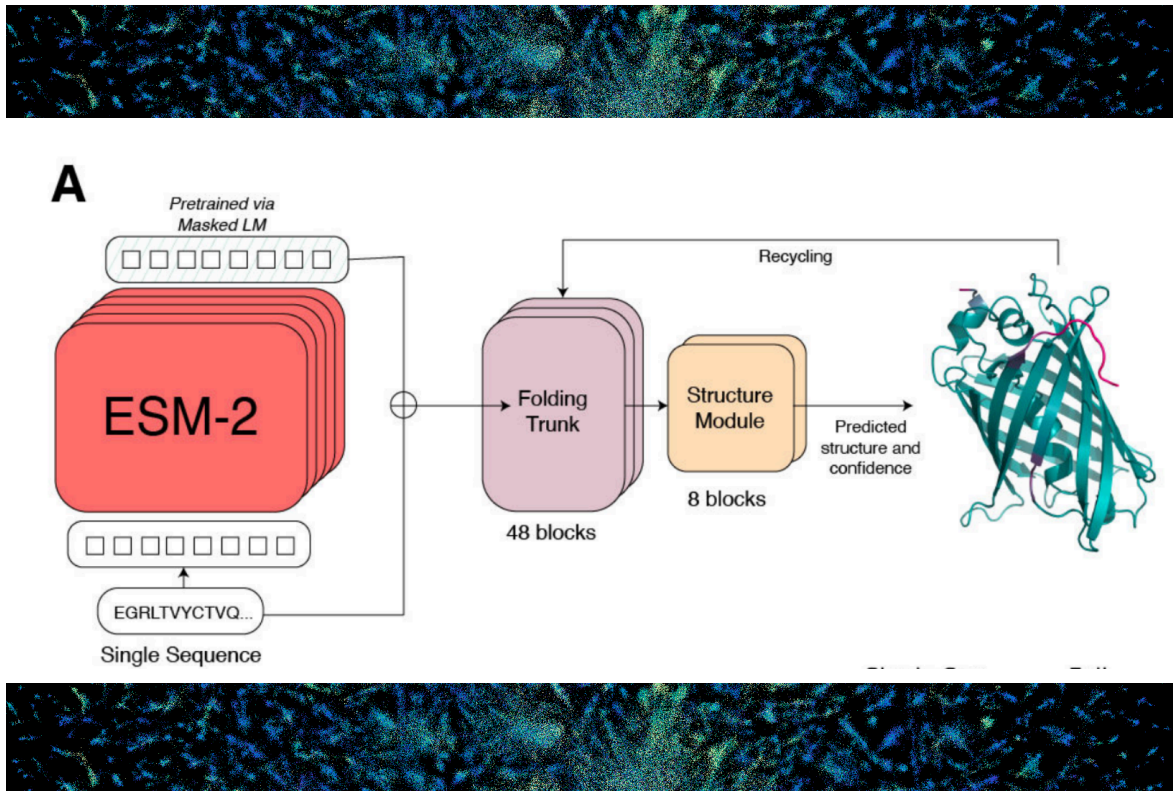
"AlphaFoldErrorPlot" y carga un archivo .pkl o .json para ver el Predicted Aligned Error (PAE) para cada residuo.

5. Cambia el esquema de color a "ColorPlotGreen" para ver los residuos con mayor confianza en verde y los residuos con menor confianza en rojo.

6. Colorea tu modelo de proteína por el valor de Local Distance Difference Test (LDDT) que se encuentra en la columna correspondiente al factor B en el archivo PDB. Los residuos coloreados en azul son los más confiables, mientras que los residuos en rojo son menos confiables.

7. Selecciona para colorear por dominios si tu proteína tiene múltiples dominios. Esto puede proporcionarte una visión adicional de la estructura de tu proteína.

B. ESMfold



0. Uso de Google Colab para ESMfold:

ESMfold es una poderosa herramienta basada en Deep Learning que nos permite predecir estructuras de proteínas con gran precisión. Para sacar el máximo provecho de esta herramienta, utilizaremos Google Colab, una plataforma en la nube que nos brinda un entorno de programación basado en Jupyter Notebook, y nos da acceso a recursos computacionales como GPUs, lo que es fundamental para el procesamiento eficiente requerido por ESMfold.

Para comenzar, accedemos al enlace proporcionado para ingresar a Google Colab: <https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/ESMFold.ipynb> U na vez en la plataforma, asegúrate de que estás conectado a una GPU. Puedes verificar esto yendo al menú de configuración y seleccionando "GPU" como tipo de acelerador. La GPU nos permitirá ejecutar nuestros modelos de Deep Learning de manera mucho más rápida y efectiva.

Una vez que estamos conectados a una GPU, estamos listos para ejecutar los códigos de ESMfold. En Google Colab, los códigos se dividen en celdas y para ejecutar cada celda, simplemente debemos hacer clic en el botón de "play" que se encuentra a la izquierda de cada una. Esto nos permitirá obtener resultados en tiempo real y realizar ajustes según sea necesario para mejorar nuestras predicciones de estructuras de proteínas.

1. Descarga e instalación de los paquetes y softwares

A lo largo de todo el proceso se descargan tanto paquetes como softwares necesarios para el análisis , evaluación y generación de los modelos de docking así como las poses:

- **ESMfold:**ESMFold es una herramienta de predicción de estructuras de proteínas que utiliza modelos de lenguaje a gran escala. Su función principal es determinar la estructura tridimensional de una proteína a nivel atómico, basándose únicamente en la secuencia individual de la proteína. A diferencia de otras herramientas similares, ESMFold destaca por su velocidad de inferencia significativamente mayor, lo que facilita el estudio de las proteínas metagenómicas en un tiempo práctico.

- **Py3DMol:** Es una biblioteca Python para visualización molecular en 3D. Nos permite explorar y comprender la estructura tridimensional de nuestras biomoléculas de interés.

- **OpenFold:** Es un software de predicción de estructuras de proteínas basado en inteligencia artificial. Una reproducción fiel pero entrenable en PyTorch del AlphaFold 2 de DeepMind. Utiliza algoritmos de vanguardia para convertir las secuencias de proteínas en modelos tridimensionales precisos. OpenFold se destaca por su capacidad para manejar una gama más amplia de secuencias, ofreciendo ventajas significativas para la investigación en proteómica y biología estructural.

3. Predicción Estructural (Colab)

jobname: " 2HIQ "

sequence: " RSYEQMETDGERQNATEIRASVGKMGIDGIGRFYIQM "

copies: 1

num_recycles: 3

sequence:

RSYEQMETDGERQNATEIRASVGKMGIDGIGRFYIQMCTELKLSDYEGRLIQNSLTIERMVL
SAFDERRNKYLEEHPSAGKDPKKTGGPIYRRVDGKWRRELILYDKEEIRRIWRQANNGD
DATAGLTHMMIWHSNLNDATYQRTRALVRTGMDPRMCSLMQGSTLPRRSGAAGAAVKG
VGTMVMELIRMIKRGINDRNFWRGENGRRTIAYERM CNILKGKFQTAQRTMVDQVRE
SRNPGNAEFEDLIFLARSALILRGSAVHKSCLPACVYGSAVASGYDFEREGYSLVGIDPFR
LLQNSQVYSLIRPNENPAHKSQLVWMACHSAFEDLRVSSFIRGTVKVPGRGLSTRGVQI
ASNENMETMESSTLELR SRYWAIRTRSGGNTNQQRASGQISIQPTFSVQRNLFPDRPTI

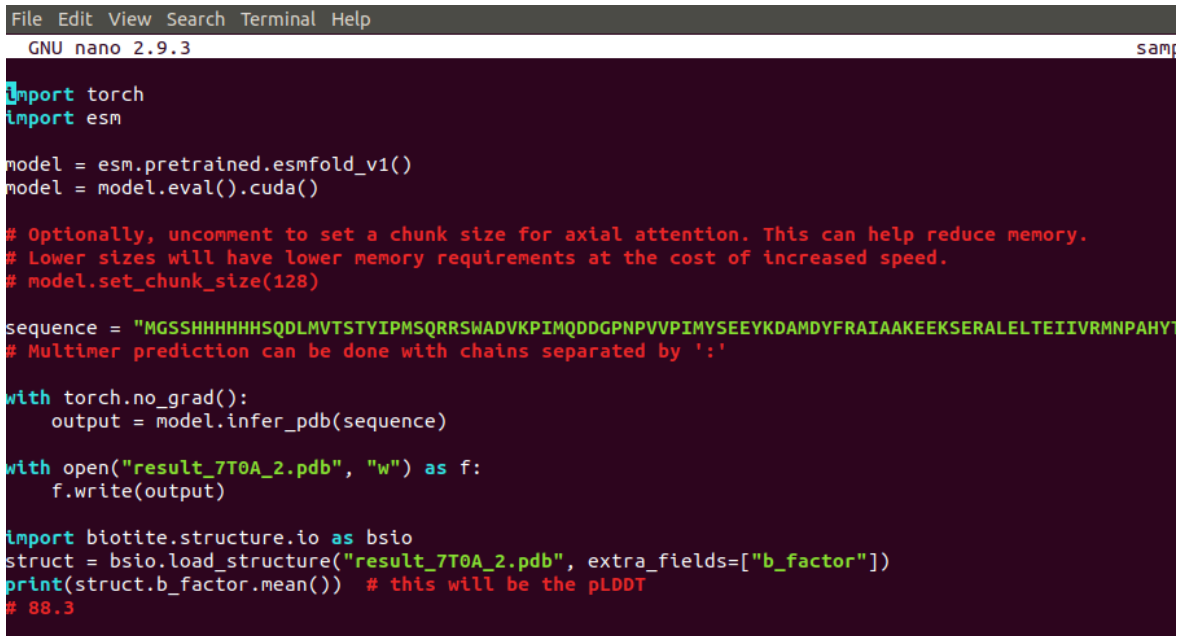
MAAFTGNTGRTSDMRTEIIRLMESARPEDVSFQGRGVFELSDEKATSPIVPSFDMSNE
GSYFFGDNA

--copies: te permite evaluar conformaciones multiméricas y heterogéneas

--num_recycles: cantidad de ciclos

4. Predicción Estructural (local)

Para poder usar el programa de forma local se usará una inferencia con ESMfold mediante torch a partir de la secuencia en un ambiente de python por lo que se deberá generar previamente un archivo .py definiendo el tamaño de chunk el cual estará directamente relacionado al gasto de memoria del GPU.



```
File Edit View Search Terminal Help
GNU nano 2.9.3 sam

import torch
import esm

model = esm.pretrained.esmfold_v1()
model = model.eval().cuda()

# Optionally, uncomment to set a chunk size for axial attention. This can help reduce memory.
# Lower sizes will have lower memory requirements at the cost of increased speed.
# model.set_chunk_size(128)

sequence = "MGSSHHHHHSQDLMTSTYIPMSQRRSWADV KPI MQDDGPNPVVPIMYSEEYKDAMDYFRAIAAKEEKSERALEL TEIIVRMNPAHYT"
# Multimer prediction can be done with chains separated by ':'

with torch.no_grad():
    output = model.infer_pdb(sequence)

with open("result_7T0A_2.pdb", "w") as f:
    f.write(output)

import biotite.structure.io as bsio
struct = bsio.load_structure("result_7T0A_2.pdb", extra_fields=["b_factor"])
print(struct.b_factor.mean()) # this will be the pLDDT
# 88.3
```

Como output se generará una media del b factor. así como un archivo pdb con la estructura esperada.

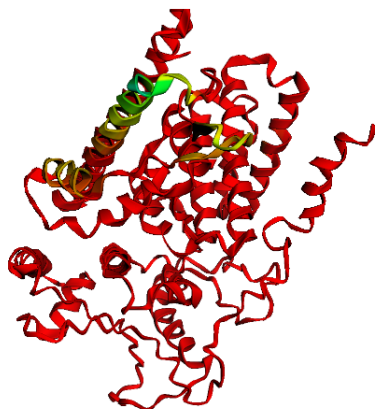
5. Visualización de resultados en colab con Py3Dmol

color: confidence

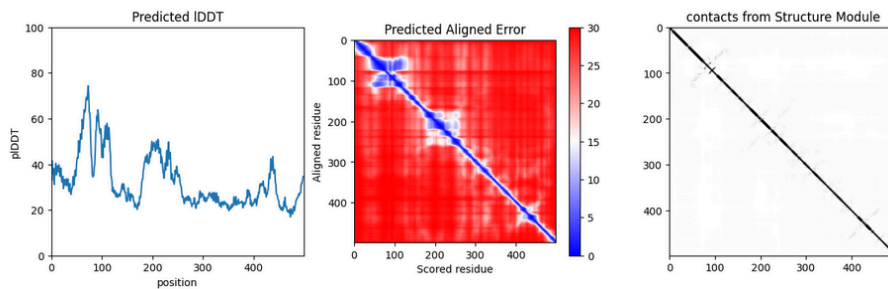
show_sidechains: ☐

show_mainchains: ☐

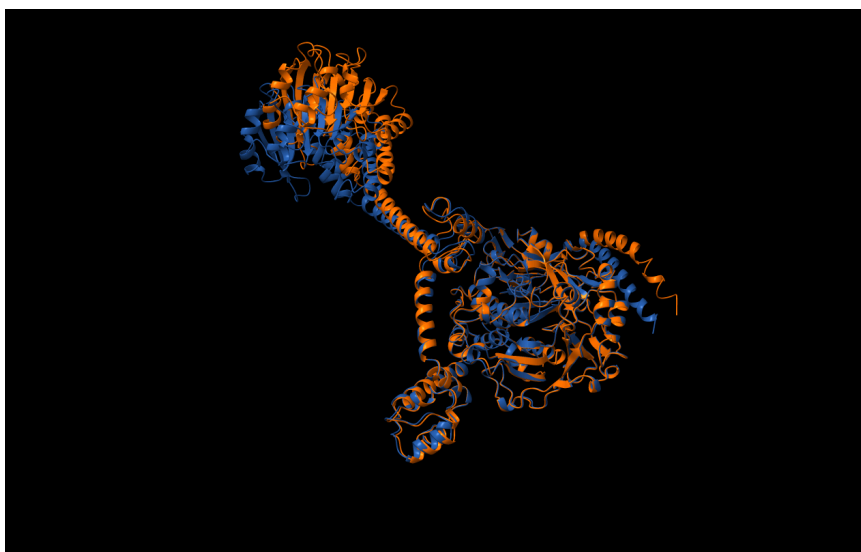
[Show code](#)



[Show code](#)



6. Visualización de resultados local con ChimeraX



Informe

1. Usando un visualizador (de preferencia chimeraX) muestre las diferencias estructurales entre el modelo estructural de 7YEU generado por alphafold2 y el modelo generado por ESMfold. En este sentido analice las métricas de confianza correspondientes y elija la estructura que tenga una mayor probabilidad de representar en realidad a la proteína.
2. Dada la velocidad con la que se pueden modelar estructuras en ESMfold ,escoja un organismo y modele las 1 proteínas a elección. (tenga en cuenta el límite de aminoácidos que se puede modelar en colab) En los casos en los que se tenga una imagen cristalográfica haga una comparacion de las estructuras en alguno de los visualizadores disponibles , en caso sea ChimeraX se podrá hacer esto con la herramienta matchmaking la cual permitirá alinear dos estructuras distintas (.pdb) y generará el RMSD; en los casos en los que no haya una imagen cristalográfica, evalúe los parámetros de calidad y concluya especificando si la estructura es lo suficientemente representativa de la proteína real de acuerdo a estos parámetros o no lo es.
3. La bioinformática, y particularmente el modelado estructural con AlphaFold, ha transformado nuestro enfoque en el estudio de proteínas que anteriormente carecían de una estructura cristalográfica determinada experimentalmente. Sin embargo, es crucial que mantengamos un enfoque crítico y no confiemos ciegamente en los modelos generados por el software, sino que debemos evaluar detenidamente sus métricas de calidad.

Con esto en mente, le pedimos que utilice ColabFold para predecir y evaluar las estructuras de dos proteínas que carecen de una estructura cristalográfica. La primera proteína es TSOL18 (código uniprot O96346), secretada por el parásito *Tenia solium*, hipotetizada para facilitar la invasión de las células epiteliales del intestino humano. La segunda es GRA8 (código uniprot A0A086L8W7), secretada por *Toxoplasma gondii*, cuya importancia se supone en el diagnóstico de la toxoplasmosis crónica.

Deberá predecir y evaluar las estructuras de estas proteínas, analizar las métricas de calidad pLDDT y PAE, y ofrecer una interpretación detallada de los resultados. Si observa diferencias en la calidad de las predicciones, le pedimos que especule sobre las posibles causas, considerando el funcionamiento y los requisitos de entrada de AlphaFold. ¿Podría la profundidad del MSA influir en la calidad de la predicción?

Tarea adicional

1. Aplique las capacidades multiméricas del software ColabFold para predecir y evaluar un complejo proteico. En este caso, estará trabajando con la proteína Btn3a3, un importante receptor humano que actúa como antiviral contra los virus de la influenza A, y la nucleoproteína NP del virus de la influenza A.

La interacción de Btn3a3 con la nucleoproteína NP de la influenza A inhibe la replicación del virus, pero se ha encontrado que ciertas mutaciones pueden afectar esta interacción, así como también diferentes linajes como el de la H1N1. Queremos que prediga y evalúe esta interacción multimérica utilizando las métricas de calidad pLDDT y PAE de ColabFold.

Además, le pedimos que reflexione sobre las siguientes cuestiones: ¿Cree que existe interacción entre estas dos proteínas? Analice el gráfico de PAE, y el PLDDT centrando su atención en la interfaz de interacción entre ambas proteínas y discuta los resultados.

2. Se modelará la proteína SmF en levadura, que forma un anillo heptamérico y es parte integral del complejo de ribonucleoproteínas nucleares pequeñas (snRNP). Este complejo juega un papel esencial en el procesamiento del ARN pre-mensajero. La estructura del anillo heptamérico de SmF está depositada en la base de datos Protein Data Bank con el código 1N9R.

Para comenzar, utiliza ColabFold (ESMfold o Alphafold) para predecir la estructura de la proteína SmF en su forma monomérica (código Uniprot P33334). Analiza la calidad de la estructura predicha con las métricas pLDDT y PAE. Secuencialmente, prediga la estructura del anillo heptamérico de SmF. Una vez que tengas tus modelos, debes evaluar y comparar la calidad de las predicciones para la forma monomérica y heptamérica. ¿Las métricas pLDDT y PAE son consistentes entre estos dos estados? ¿Cómo afecta la oligomerización a la conformación de la proteína y la calidad de la predicción?

Finalmente, compara tus predicciones con la estructura experimental 1N9R. Si hay diferencias, ¿en qué partes de la proteína se encuentran y qué podrían significar estas diferencias? Especula sobre las posibles razones de cualquier discrepancia entre tus predicciones y la estructura experimental. ¿Podría la profundidad de la MSA o el hecho de que SmF normalmente no es una proteína monomérica influir en la precisión de la predicción?