

Conotoxinas

Onna Nayyu Leyva Alcantara

2023-05-09

Reporte

Introducción

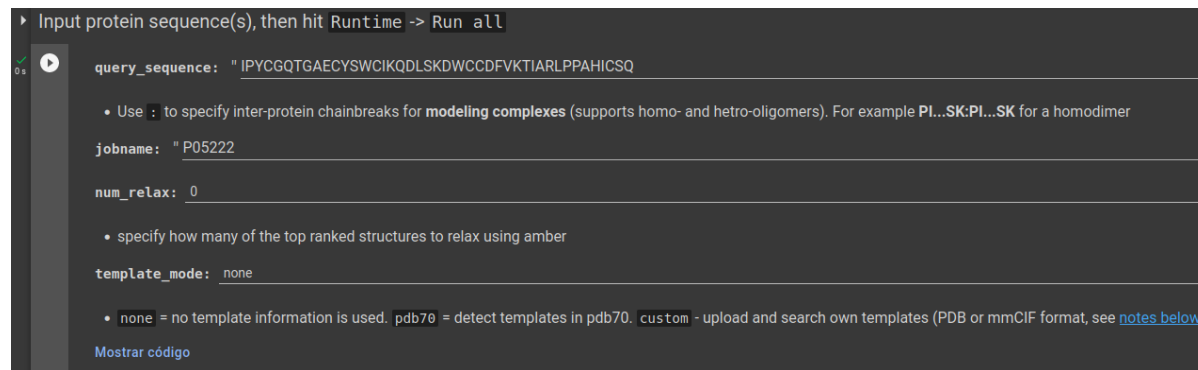
Las conotoxinas o conopeptidos son péptidos neurotóxicos producidos por aproximadamente 800 especies de caracoles marinos del género *Conus*, cada uno produce entre 100 y 200 de estos. Estos son pequeños en comparación con los venenos de víboras o arañas, compuestos por entre 10-35 residuos.

Dada su alta especificidad se han utilizado para determinar la selectividad de determinados fármacos a canales iónicos. También han sido utilizados en la síntesis de analgésicos potentes que bloquean la liberación de los neuro-transmisores del dolor, previniendo su propagación al cerebro.

Objetivo

Realizar una predicción estructural utilizando *AlphaFold2* y *RoseTTAFold* a partir de secuencias primarias de cinco diferentes superfamilias de conotoxinas (una de cada familia).

Descripción de los bloques de código

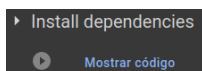


```
▶ Input protein sequence(s), then hit Runtime -> Run all
0s
query_sequence: "IPYCGQTGAECYSWCIKQDLSKDWCCDFVKTIARLPPAHICSQ"
  • Use : to specify inter-protein chainbreaks for modeling complexes (supports homo- and hetero-oligomers). For example PI...SK:PI...SK for a homodimer
jobname: "P05222"
num_relax: 0
  • specify how many of the top ranked structures to relax using amber
template_mode: none
  • none = no template information is used, pdb70 = detect templates in pdb70, custom = upload and search own templates (PDB or mmCIF format, see notes below)
Mostrar código
```

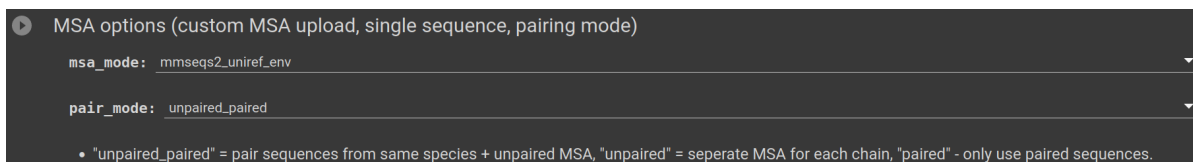
En este primer bloque de código de AlphaFold2 podemos apreciar que los primeros parámetros a ingresar es la secuencia query de la proteína (secuencia de aminoácidos) y posteriormente el nombre del trabajo que en este caso se colocó el ID de la proteína.

Posteriormente *Num_relax* (número de relajaciones adicionales) se refiere a ajustar la estructura de la proteína de manera que minimice la energía potencial de la molécula, lo que lleva a una conformación más estable y realista. En este caso no se llevará a cabo ninguna relajación adicional al terminar la predicción.

Por último, *template_mode* controla cómo se utilizan las estructuras de proteínas conocidas (también conocidas como plantillas o templates). En este caso no se usarán plantillas.



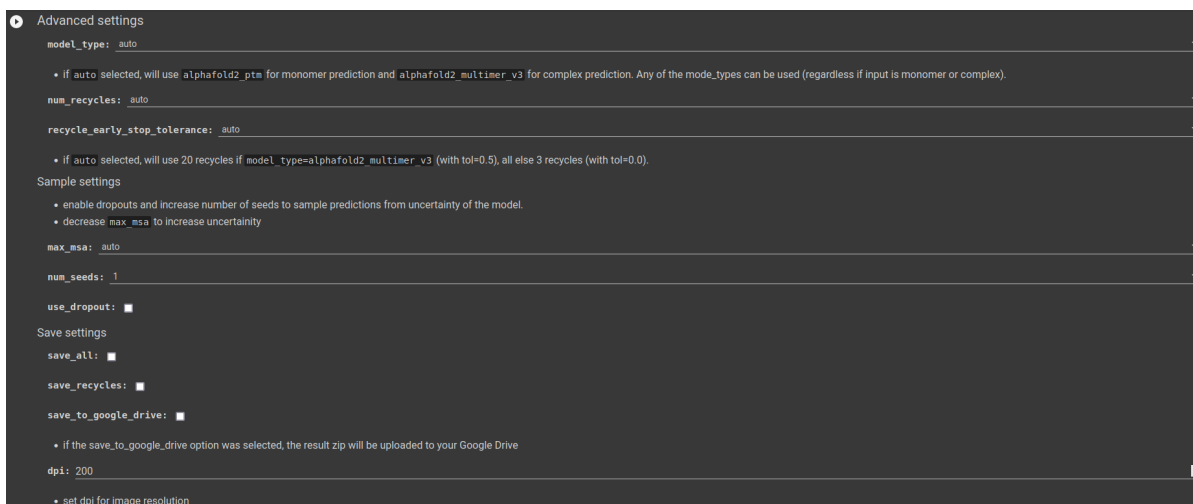
En este bloque se instalan los paquetes y bibliotecas que necesitarán los siguientes bloques de código para funcionar correctamente.



En dicho bloque se encuentran las opciones para el alineamiento de secuencias múltiples (msa).

Primeramente tenemos *msa_mode* (modo de msa), en este caso se elige *mmseqs2_uniref_env*, “mmseqs2” es programa que busca secuencias homólogas, “uniref” proporciona un conjunto de secuencias de referencia no redundantes, por lo tanto, esta opciones busca secuencias homólogas en un conjunto no redundante de secuencias.

Por otro lado *pair_mode* (modo de pareado), hace referencia a la forma de emparejamiento de las secuencias para el alineamiento múltiple, en este caso *unpaired_paired* combina secuencias emparejadas de la misma especie con secuencias no emparejadas de diferentes especies.

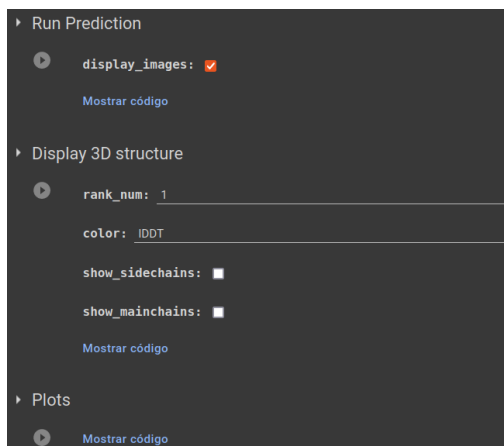


En primer lugar tenemos el *model_type* (tipo de modelo) en este caso se especifica el *auto* el cual corre automáticamente todos los tipos de modelo y selecciona el proporciona la mejor predicción para la proteína dada.

De manera similar especificar *auto* provoca que AlphaFold2 determine dependiendo de la proteína el número optimo de iteraciones tanto del *num_recycles* que implica tomar la estructura predicha, volver a introducirla en el modelo y ejecutar una nueva predicción, como del *recycle_early_stop_tolerance* que indica que la proteína predicha se ha estabilizado y ya no mejora significativamente. De la misma manera en *max_msa* (número máximo de secuencias que se incluirán en el msa) *auto* realiza la misma función.

Por otra lado, *num_seeds* especifica el número de secuencias de semillas que se utilizarán para generar el MSA. El valor predeterminado es 1. *use_dropout* indica el uso o no de la tecnica de dropout (previene el sobreajuste). En este caso no.

Por último, se tienen las opciones de guardar todos los pasos intermedios o no, guardar los reciclados o no y si guardar estos en la nube o no. Además se puede especificar la resolución de la imagen.



Para finalizar se corre la predicción, puedes elegir que se muestren o no las imágenes. En este caso si queremos que se muestren.

Más adelante corres el código para que te muestre la estructura 3D, en este caso se elige 1 en *rank_num* por lo que se mostrara solo la estructura de la proteína solo con la puntuación más alta. Después se selecciona la escala de color de calidad de la estructura *IDDT*. Al final escoges si se mostrarán las cadenas secundarias, primarias, ambas o ninguna.

Por último, corres el código para hacer gráficos que posteriormente se explicaran las obtenidas para cada proteína.

Destacar que para RoseTTAFold también se usan bloques de código, pero estos son muy similares a los de AlphaFold2.

Predicción estructural de conotoxinas

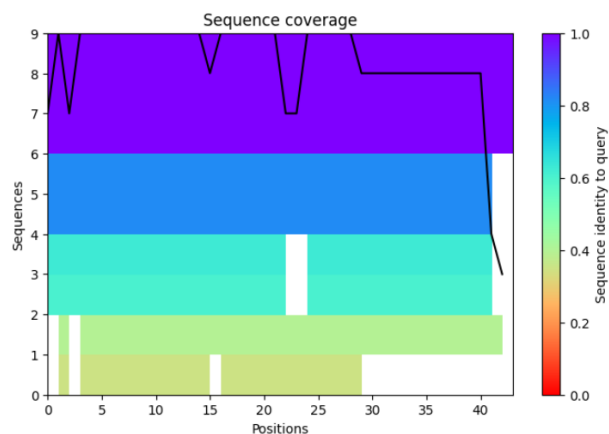
1-

Familia K ID de la proteína (jobname): **P05222**

Secuencia query: **IPYCGQTGAECYSWCIKQDLKDWCCDFVKTIARLPPAHICSQ**

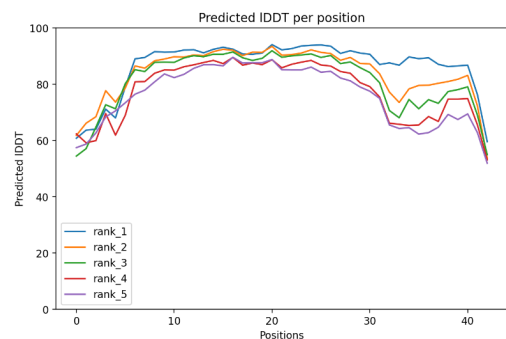
Modelo con AlphaFold2

Al correr la predicción obtenemos la siguiente gráfica que nos indica la cobertura que tuvo la secuencia y al mismo tiempo representa el número de MSA que el programa toma. Esto lo enfatizo, ya que el tamaño MSA para todas los modelos se específico como **AUTO** (anteriormente explicado), y en todo el reporte no mencionare el MSA, ya que verlo como cobertura es más visual para mi, además el específico MSA de cada modelo lo añadire en el excel.

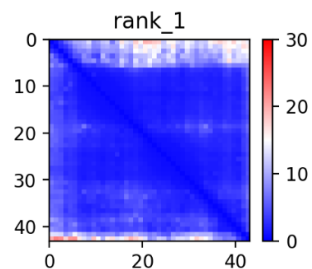


Podemos notar que la cobertura por posición es medianamente buena, ya que se mantiene por encima de siete secuencias, sin embargo podemos notar al final un descenso en la cobertura.

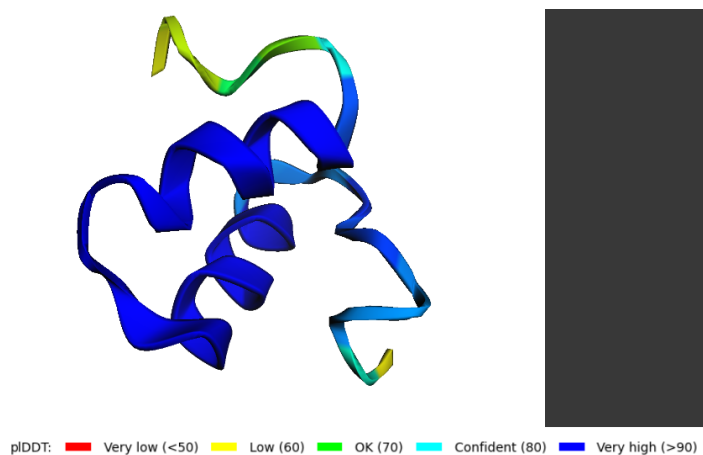
También se despliegan los mejores cinco modelos, pero como se observa en la imagen el mejor es el que analizaremos (azul). De la misma manera solo nos fijaremos en la mejor para las siguientes proteínas.



También obtenemos una matriz de distancias que ilustra la manera de la que la proteína está plegada a grandes rasgos, por ejemplo, en este caso indica que los últimos aminoácidos están lejos de los primeros, mientras que todos los demás están bastante cerca.



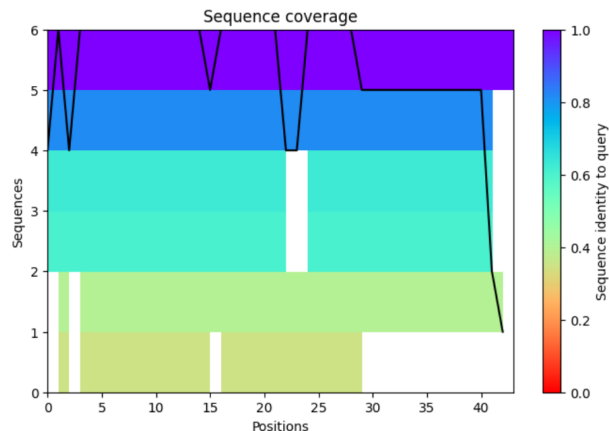
Aunado con la anterior podemos ver como en la siguiente estructura 3D concuerda con la matriz de distancias.



Esta cuenta con un **pLDDT** en promedio de **86.4** que se puede entender como la confianza en la precisión del modelo. En este caso podemos notar que a mitad de la proteína la confiabilidad es un poco baja con 80 de pLDDT y en los extremos esta baja a 60.

Modelo con RoseTTAFold

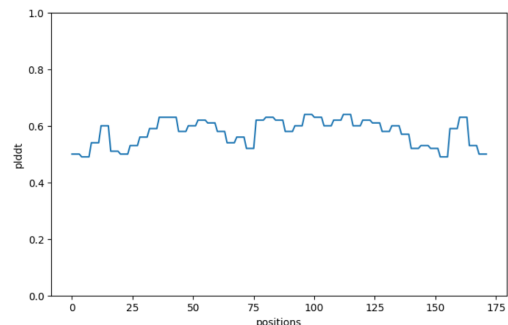
Al correr el código, primeramente se obtiene la siguiente gráfica de cobertura.



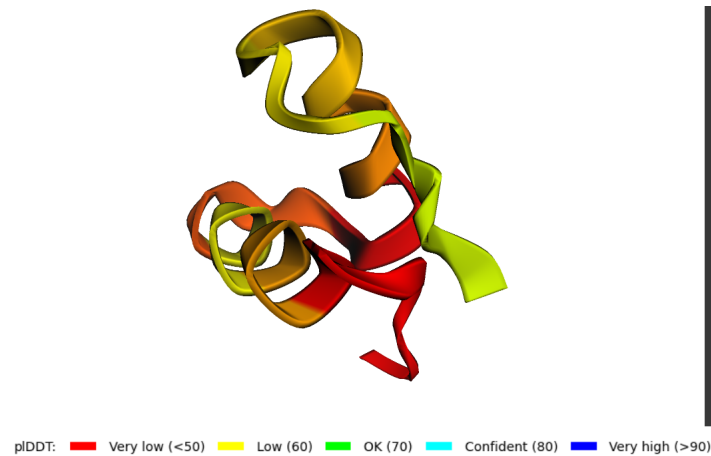
En comparación con AlphaFold2 se observa una cobertura casi idéntica a lo largo de la proteína pero en vez de que la máxima sea de 9 es de 6, por lo que, aunque el patrón es similar la cobertura por residuo es menor.

En relación con lo anterior el **pLDDT** en promedio de la predicción es de **0.57**. Notar que el valor de pLDDT está dividido entre cien, por lo que está en porcentaje a diferencia de AlphaFold.

Cabe aclarar que posteriormente no colocare la siguiente gráfica dado que se puede obtener información similar con la estructura en 3D.



En complemento con lo anterior la estructura en 3D se ve de la siguiente forma.



Por lo tanto, la confianza en la precisión del modelo es baja. En conclusión, la predicción de AlphaFold2 es buena en general, además de ser mejor que la de RoseTTAFold.

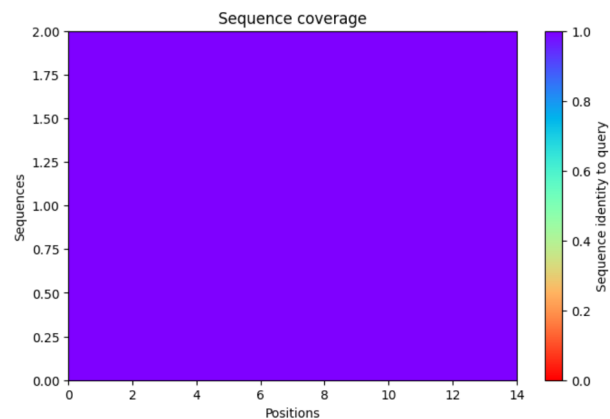
2-

Familia T ID de la proteína: **P06123**

Secuencia query: **VADDCCVGKVGTTCC**

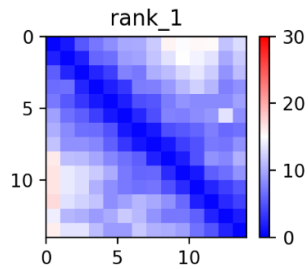
Modelo de AlphaFold2

En primer lugar obtenemos la siguiente gráfica.



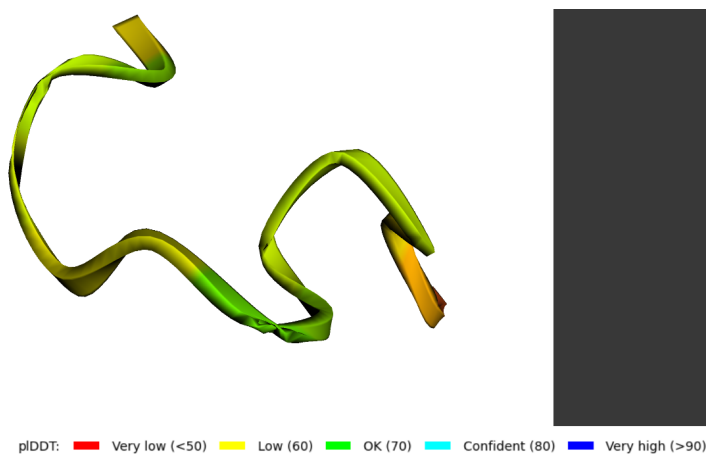
Podemos observar que la cobertura por posición es buena, sin embargo es importante notar que solo hay un máximo de dos.

Realizamos la matriz de distancias.



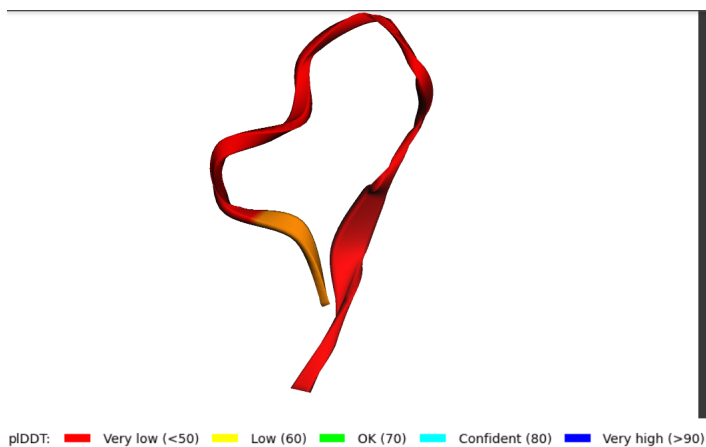
Observamos que la proteína se pliega de manera que sus extremos quedan separados y los aminoácidos intermedios no se alejan mucho entre ellos.

Posteriormente se obtiene la estructura 3D, donde se observa un plegamiento similar al descrito, además, la confianza en el modelo de predicción o el **pLDDT** está entre **60-70** mayormente, pero en promedio está en 61.3, por lo tanto, se considera como bajo.



Modelo de RoseTTAFold

Al correr el código para la gráfica de *sequence coverage* arrojo como resultado que solo se encontro una secuencia. Al realizar la gráfica del **pLDDT** por posición se obtuvo un promedio de **0.46**, por lo que, al realizar la estructura 3D se obtuvo la siguiente imagen.



Se puede observar que la confianza de la predicción del modelo es muy baja. En comparación con AlphaFold2 este cuenta con muchas menos confianza, sin embargo, ambos modelos no se podrían considerar modelos confiables.

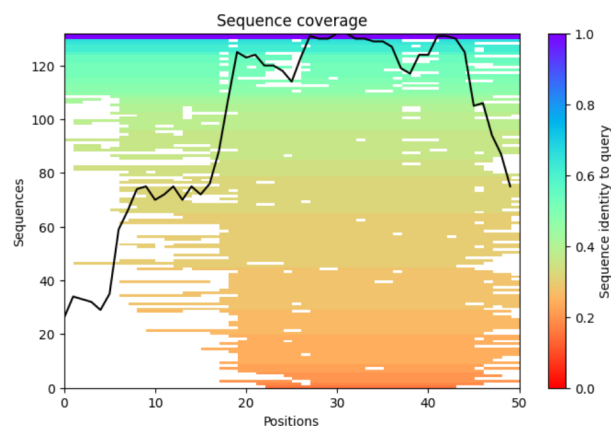
3-

Familia B2 ID de la proteína: **P07081**

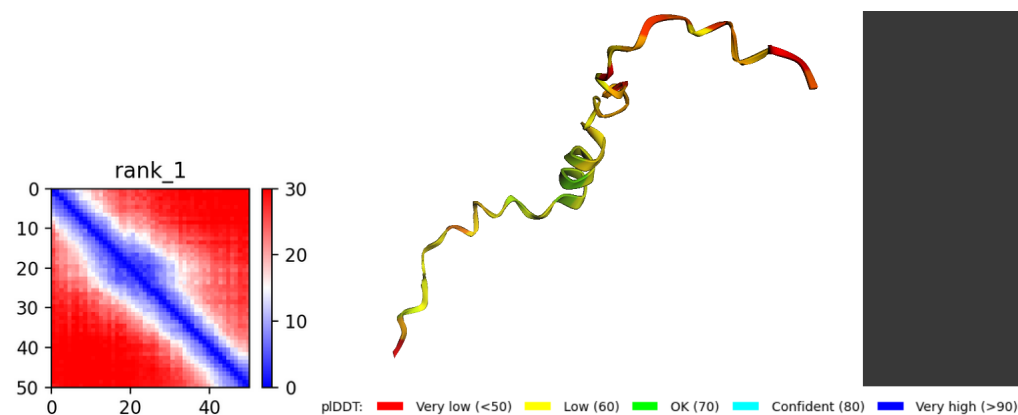
Secuencia query: **FPQRRDGAPAENLKSFDPAQMAGGMPNMQGMQPMGNIGPRP-NAAFQP**

Modelo de AlphaFold2

En la primera gráfica notamos como los primeros 20 aminoácidos tienen una cobertura menor a 80, que en comparación con los siguientes aminoácidos es menor, ya que estos tienen 110 secuencias de cobertura aproximadamente.



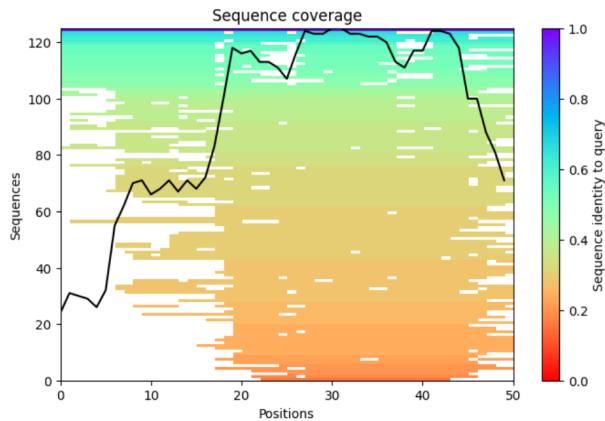
La siguiente gráfica describe a una proteína, donde tanto los extremos como las partes intermedias de la proteína están alejadas entre sí completamente, pudiendo dar como resultado una proteína lineal.



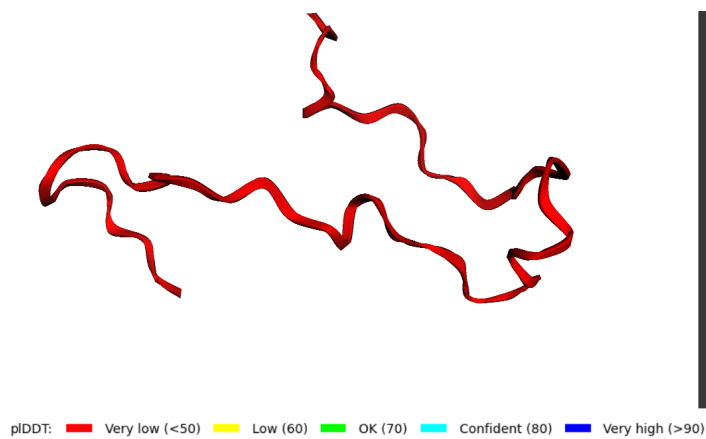
En relación a la matriz de distancias podemos observar que la estructura 3D describe una proteína lineal en su mayoría. Por otro lado, el **pLDDT** es en promedio **57.2**, lo cual indica una confiabilidad baja en el modelo predicho.

Modelo de RoseTTAFold

Al obtener la gráfica de "sequence coverage" de RoseTTAFold notamos que es idéntica a la elaborada por AlphaFold2. De esto podemos decir dos cosas, la información con la que cuenta cada programa de estas secuencias en específico es la misma; por otro lado, ambos utilizan algoritmos similares para realizar el análisis de cobertura.



Posteriormente se obtiene un **pLDDT** de **0.29** en promedio y se elabora la estructura 3D siguiente.



Se observa como a lo largo de toda la proteína el **pLDDT** es muy bajo, indicando la poca confianza hacia este modelo. En comparación con el modelo predicho por AlphaFold2 este es menos confiable, sin embargo, ambos tienen muy baja confiabilidad.

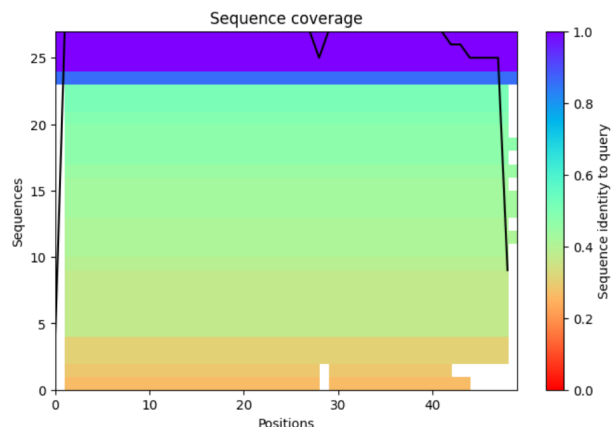
4-

Familia F ID de la proteína: **P09942**

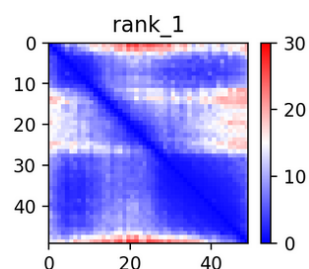
Secuencia query: **AMHSCAIVNDYDDRRWSSYNVAEFKDRSLFRTMVTDLQGCLNYF-FQIRP**

Modelo de AlphaFold2

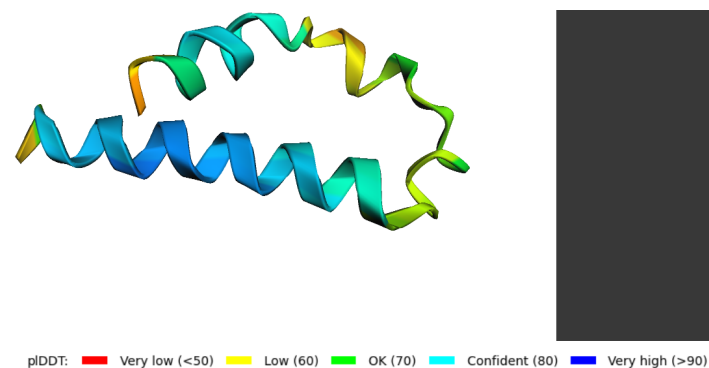
Se puede observar que la cobertura de la siguiente gráfica para casi todos los aminoácidos es buena.



En la siguiente matriz notamos que los aminoácidos intermedios están alejados de los demás, los primeros y últimos tienen cierta cercanía.

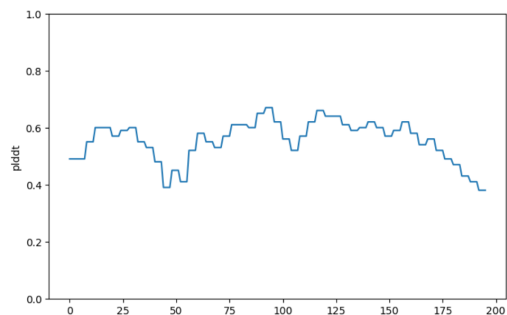


Al hacer la estructura 3D observamos que la confianza es aceptable (no confiable) en general con un **pLDDT** en promedio de **74**, aun contando con una cobertura por aminoácidos alta, esto podría deberse a que se está tratando de predecir una proteína con bucles.

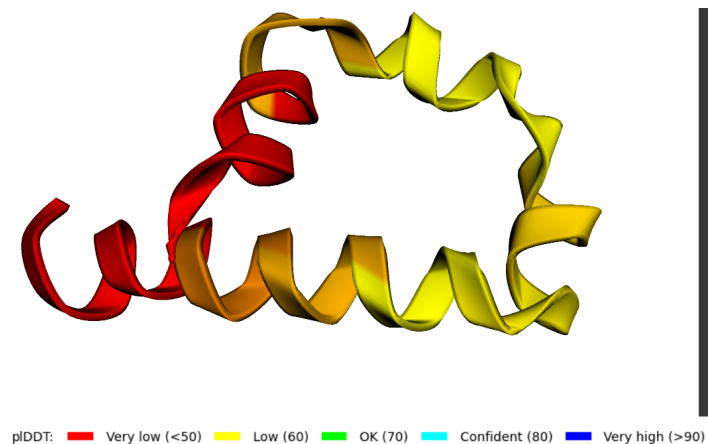


Modelo de RoseTTAFold

Dado que la gráfica de "sequence coverage" se ve muy similar a la que AlphaFold2 crea no la incluí. En la siguiente podemos observar que el **pLDDT** se mantiene entre **0.4-0.6**, más específicamente en promedio tiene un valor de **0.55**, por lo que, su confiabilidad es baja.



La estructura predicha es similar a la de AlphaFold2 y de hecho las partes en las que difieren son en las que la confiabilidad de ambos modelos es baja. Destacar que ambas predicciones cuentan con los bucles.



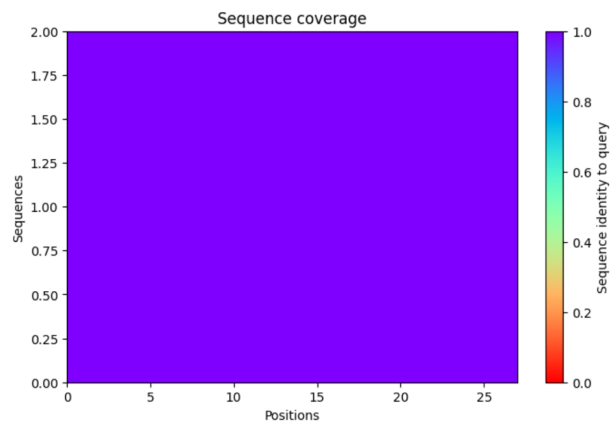
5-

Familia H ID de la proteína: **P05500**

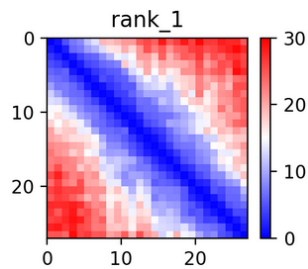
Secuencia query: **GCGMMRVTVQQPLSPEALSWTPNCNVS**

Modelo de AlphaFold2

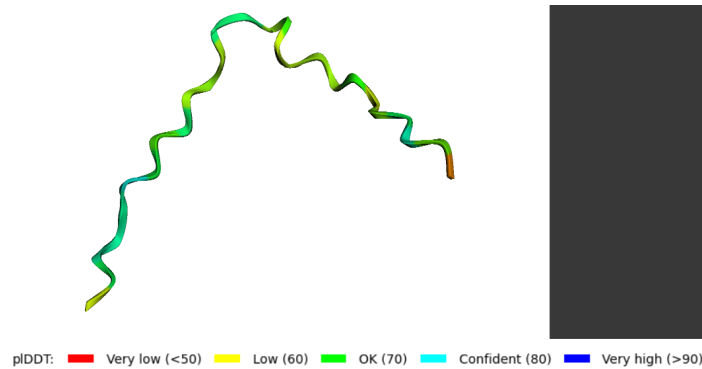
Aunque la cobertura es alta solo se encontraron dos secuencias.



En el siguiente gráfico notamos que los aminoácidos entre sí se encuentran alejados, en otras palabras podríamos pensar que la estructura será muy lineal.

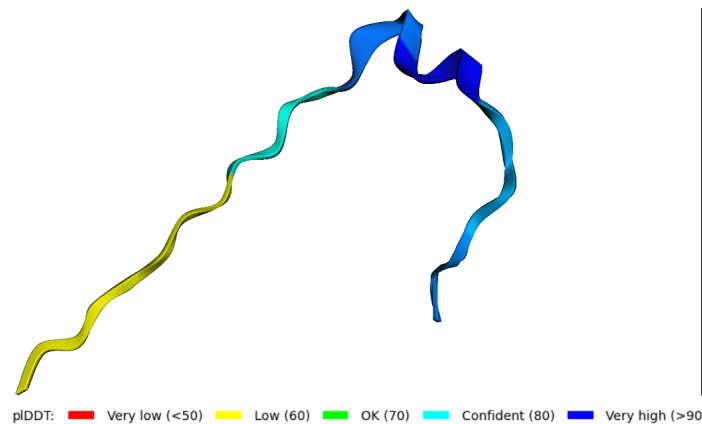


Obtenemos un **pLDDT** en promedio de **68.6** el cual es aceptable, sin embargo, la confianza del modelo no es buena.



Modelo de RoseTTAFold

Al correr el código para la gráfica de “sequence coverage” nos indica que solo se encontro una secuencia. Posteriormente obtenemos el **pLDDT** en promedio de **0.55** el cual es bajo, aun tomando en cuenta que tiene zonas con muy alta confiabilidad. Por lo tanto, de igual manera a la predicción de AlphaFold2, ninguna es confiable.



Conclusión

Es importante destacar que la conotoxina perteneciente a la familia “K” tuvo un valor “pLDDT” bastante alto con AlphaFold2, aun cuando para RoseTTAFold no fue muy confiable. Esto podría deberse a que AlphaFold2 cuenta con mayor cantidad de secuencias en este caso.

En conclusión, la predicción de modelos para las conotoxinas expuestas tienen poca o aceptable confiabilidad, debido a dos cosas; la baja cantidad de secuencias de cobertura con las se cuentan, aun cuando el

predeterminado de ambos programas utiliza la cantidad que proporcione la mejor predicción; y la mayoría conotoxinas tiene varios bucles, dificultando así su predicción.

Bibliografía

- Wikipedia contributors. (n.d.). Conotoxina. Wikipedia, The Free Encyclopedia. <https://es.wikipedia.org/w/index.php?title=Conotoxina&oldid=120120164>