

Determinación de la estructura tridimensional de conotoxinas de la superfamilia A, O1 aisladas de caracol “*Conus ermineus*” por secuenciación con Illumina utilizando el software AlphaFold2

María Camila Peñaloza Granados, Marla Yaneira Sánchez Alvarez

Trabajo de Grado para optar al título de Química

Director

Hermínul De Jesús Cano Calle

Doctor en Química

Codirector

Enrique Mejía Ospino

Doctor en Química

Universidad Industrial de Santander

Facultad de Ciencias Básicas

Escuela de Química

Química

Bucaramanga

2023

### **Dedicatoria**

A nuestras familias por haber sido apoyo durante estos años de proceso, por su paciencia, comprensión, fuerza y amor. Por enseñarnos a no rendirnos en los momentos difíciles y por guiarnos con sus sabios consejos.

### **Agradecimientos**

A la Universidad Industrial de Santander de la cual nos sentimos profundamente orgullosas, por formarnos, enriquecernos en conocimientos, brindarnos oportunidades y enseñanzas durante este largo camino. Gracias a todas las personas que fueron partícipes de este proceso. Gracias a nuestros directores por ser la guía de nuestro proyecto y principalmente a nuestra familia que son el motor de impulso de todos nuestros sueños.

## Tabla de Contenido

Introducción	13
1 Objetivos	20
1.1 Objetivo general	20
1.2 Objetivos específicos	20
2 Métodos	21
3 Resultados	24
3.1 AlphaFold2	24
3.2 PEP-FOLD2	28
3.3 Comparación	32
4 Discusión	38
4.1 AlphaFold2	38
4.2 PEP-FOLD2	40
4.3 Comparación	40
5 Conclusiones	44
6 Recomendaciones	45
Referencias Bibliográficas	45

**Lista de Figuras**

Figura 1	Estructuras tridimensionales de la superfamilia A obtenidas mediante AlphaFold2	25
Figura 2	Estructuras tridimensionales de la superfamilia O obtenidas mediante AlphaFold2	27
Figura 3	Estructura tridimensional del péptido Alpha-conotoxin MII obtenida mediante AlphaFold2	28
Figura 4	Estructuras tridimensionales de la superfamilia A obtenidas mediante PEP-FOLD2	29
Figura 5	Estructuras tridimensionales de la superfamilia O obtenidas mediante PEP-FOLD2	31
Figura 6	Estructuras tridimensionales obtenidas de la comparación entre AlphaFold2 y técnicas experimentales (RMN y/o RX) para la superfamilia A	33
Figura 7	Estructuras tridimensionales obtenidas de la comparación entre AlphaFold2 y técnicas experimentales (RMN y/o RX) para la superfamilia O	34
Figura 8	Estructura tridimensional de la conotoxina MII obtenida de la comparación entre AlphaFold2 y RMN	37
Figura 9	Comparación de las conotoxina O1 y MVII	42
Figura 10	Comparación de las conotoxina O1 y SO-3	42

**Lista de Tablas**

Tabla 1	Alineación del marco de cisteínas tipo I	14
Tabla 2	Alineación del marco de cisteínas tipo IV	15
Tabla 3	Alineación del marco de cisteínas tipo VI/VII	15

## Lista de Apéndices

(Ver apéndices adjuntos en el CD y pueden visualizarlos en la Base de Datos de la Biblioteca UIS)

**Apéndice A** Secuencias de conotoxinas del caracol *Conus ermineus* obtenidas de Illumina

**Apéndice B** Valores de pLDDT de cada modelo recuperados mediante AlphaFold2 para las diferentes estructuras de conotoxinas

**Apéndice C** Secuencias de conotoxinas obtenidas del Protein Data Bank a usar para la comparación con las estructuras recuperadas mediante AlphaFold2

**Apéndice D** Datos obtenidos de la comparación entre las conotoxinas obtenidas mediante AlphaFold2 y métodos experimentales

## Glosario

**AMBER** serie de campos de fuerza de la mecánica molecular clásica, desarrollada principalmente para modelar biomoléculas (Salomon-Ferrer, Case, y Walker, 2013).

**Conoserver** es una base de datos diseñada para organizar información sobre conopeptidos para un acceso fácil y conveniente al descubrimiento de conopeptidos, datos de estructura y actividad, y datos de evolución del veneno. (Kaas, Yu, Jin, Dutertre, y Craik, 2012).

**Conotoxinas** grupo de péptidos neurotóxicos ricos en cisteína aislados del veneno de los caracoles marinos del género *Conus* (Dao y cols., 2017).

**ExPASy** “servidor de bioinformática que brinda acceso a bases de datos científicas y herramientas de software en una amplia gama de disciplinas de las ciencias de la vida, incluidas proteómica, genómica, filogenia, biología de sistemas, genética de poblaciones y transcriptómica” (Artimo y cols., 2012).

**FASTA** es un formato comúnmente utilizado para almacenar datos de secuencias de nucleótidos o proteínas (Shen, Le, Li, y Hu, 2016).

**Identidad** número de caracteres que coinciden exactamente entre dos secuencias diferentes (Udayangani, 2019).

**Illumina** la tecnología de Illumina se utiliza para secuenciar el ADN y el ARN, analizar la varia-

ción genética y comprender los patrones de expresión génica. La tecnología de secuenciación de Illumina se puede utilizar para analizar el transcriptoma de un organismo (Bentley y cols., 2008). Esto implica convertir las transcripciones de ARN en ADN complementario (ADNc) mediante la transcripción inversa y, a continuación, secuenciar el ADNc mediante la secuenciación de Illumina (Kukurba y Montgomery, 2015).

**NOC3.01** es un explorador molecular gratuito para la visualización, validación y análisis de estructuras de proteínas (Cang, Chen, y Nymeyer, 2006).

**PDB** el “Protein Data Bank” es el banco de información estructural de proteínas a nivel atómico, que almacena todas las estructuras tridimensionales de las macromoléculas biológicas determinadas por la comunidad científica mundial (Rudiño-Piñera, Quintero-Hernández, y Juárez-González, 2022).

**RMSD** se usa a menudo en el modelado molecular para comparar las estructuras de dos moléculas o para evaluar la precisión de un método de modelado molecular (Sargsyan, Grauffel, y Lim, 2017).

**Similaridad** es la semejanza entre dos secuencias cuando se comparan. La similaridad representa la medida en que los residuos están alineados. Por lo tanto, secuencias similares contienen propiedades similares (Udayangani, 2019).

**Superfamilia** clasificación que reciben las conotoxinas según la homología de su secuencia señal conservada (Akondi y cols., 2014).

**SuperPose** el servidor web SuperPose utiliza un enfoque mejorado de valores propios de cuater-

niones para calcular de forma rápida y sólida superposiciones estructurales de proteínas múltiples y por pares. Cuenta con la capacidad única de superponer estructuras que varían en secuencia, tamaño o forma (Maiti, Van Domselaar, Zhang, y Wishart, 2004).

**YASARA** es un software de visualización y modelado molecular que se utiliza para simular y analizar macromoléculas biológicas como proteínas, ADN y ARN (Krieger y Vriend, 2014).

## Resumen

**Título:** Determinación de la estructura tridimensional de conotoxinas de la superfamilia A, O1 aisladas del caracol “*Conus ermineus*” por secuenciación con Illumina utilizando el software AlphaFold2 \*

**Autor:** María Camila Peñaloza Granados, Marla Yaneira Sánchez Alvarez \*\*

**Palabras Clave:** Conotoxinas, péptidos, AlphaFold2, estructura tridimensional, PEP-FOLD2.

**Descripción:** Las conotoxinas son pequeños péptidos ricos en disulfuro aislados del veneno de caracoles marinos del género *Conus*, que al unirse a canales iónicos estos péptidos poseen importantes funciones biológicas; es por ello, que se han convertido en objeto de importantes estudios, sin embargo, para comprender su función es fundamental su caracterización estructural, la cual aún se encuentra muy rezagada. Dado lo anterior, en este estudio se propuso evaluar la capacidad de reconstrucción de las estructuras tridimensionales de algunos péptidos seleccionados usando AlphaFold2, un software basado en inteligencia artificial. Igualmente, se empleó PEP-FOLD2 el cual maneja un acercamiento con campos de fuerza para reconstruir la estructura tridimensional de las mismas secuencias. Finalmente, las estructuras obtenidas con AlphaFold2 se compararon con estructuras determinadas mediante RMN y RX. En este estudio AlphaFold2 demostró que logra recuperar la estructura de la cadena principal de péptidos con valores de pLDDT altos, lo que indica una buena precisión, pero presentó deficiencias en la recuperación de regiones flexibles y no logró recuperar en su totalidad los puentes disulfuro. Por otra parte, PEP-FOLD2 no logró recuperar ningún puente disulfuro, por lo tanto, no se tuvieron en cuenta las estructuras obtenidas con este software. AlphaFold2 demostró que a pesar de no recuperar con la misma precisión todas las estructuras es más confiable que PEP-FOLD2, sin embargo, dichas diferencias en la precisión podrían deberse a que el software no cuenta con suficientes bases de datos relacionadas a péptidos, dado lo anterior, se sugiere continuar entrenando el software en la predicción de estructuras tridimensionales de péptidos para de esta manera ampliar su base de datos.

---

\* Trabajo de grado

\*\* Facultad de Ciencias. Escuela de Química. Química. Director: Herminul De Jesús Cano Calle. Doctor en química. Codirector: Enrique Mejía Ospino. Doctor en química.

## Abstract

**Title:** Determination of the three-dimensional structure of conotoxins of the A, O1 superfamily isolated from the “*Conus ermineus*” snail by Illumina sequencing using AlphaFold2 software \*

**Author:** María Camila Peñaloza Granados, Marla Yaneira Sánchez Alvarez \*\*

**Keywords:** Conotoxins, peptides, AlphaFold2, three-dimensional structure, PEP-FOLD2.

**Description:** Conotoxins are small disulfide-rich peptides isolated from the venom of marine snails of the genus *Conus*, which by binding to ion channels these peptides possess important biological functions, which is why they have become the subject of important studies, however, to understand their function it is essential their structural characterization, which is still far behind. Therefore, in this study we proposed to evaluate the ability to reconstruct the three-dimensional structures of some selected peptides using AlphaFold2, a software based on artificial intelligence. Likewise, PEP-FOLD2, which uses a force field approach to reconstruct the three-dimensional structure of the same sequences, was used. Finally, the structures obtained with AlphaFold2 were compared with structures determined by NMR and XR. In this study AlphaFold2 showed that it is able to recover the main chain structure of peptides with high pLDDT values, which indicates good accuracy, but it presented deficiencies in the recovery of flexible regions and failed to fully recover the disulfide bridges. On the other hand, PEP-FOLD2 failed to recover any disulfide bridges, therefore, structures obtained with this software were not taken into account. AlphaFold2 showed that despite not recovering all structures with the same accuracy it is more reliable than PEP-FOLD2, however, these differences in accuracy could be due to the fact that the software does not have enough peptide related databases, given the above, it is suggested to continue training the software in the prediction of three-dimensional structures of peptides in order to expand its database.

---

\* Degree Work

\*\* Faculty of sciences. School of chemistry. Chemistry. Director: Hermisul De Jesús Cano Calle. PhD in chemistry. Co-director: Enrique Mejía Ospino. PhD in chemistry.

## Introducción

Los péptidos poseen una variedad de funciones específicas debido al carácter individual de cada aminoácido y se refuerza al combinar varios aminoácidos en moléculas más grandes (Bodanszky, 1988). De esta manera, los péptidos pueden actuar como mensajeros químicos, neurotransmisores, así como estimuladores e inhibidores altamente específicos, regulando diversos procesos vitales (Bodanszky, 1988).

En el presente estudio los péptidos de principal interés son los péptidos del veneno de los caracoles marinos del género *Conus*, específicamente la especie *Conus ermineus*, los cuales son conocidos como conotoxinas, quienes son ricas en disulfuros, y actúan sobre una amplia gama de objetivos como canales iónicos, receptores acoplados a proteínas G (GPCR), transportadores y enzimas (Jin y cols., 2019). Estas conotoxinas son actualmente motivo de interés para los neurocientíficos y desarrolladores de fármacos debido a su potencia y selectividad específica no solo contra presas sino también contra objetivos de mamíferos, lo que proporciona una rica fuente de fármacos con fines terapéuticos (Jin y cols., 2019).

Estas conotoxinas se agrupan en superfamilias de genes según la similitud entre las secuencias señal. Aunque las conotoxinas en una superfamilia a menudo comparten secuencias de péptidos señal similares, los péptidos de veneno codificados tienen una diversidad estructural y funcional notable (Jin y cols., 2019). En la actualidad se reconocen 28 familias catalogadas en el ConoServer (Kaas, Westermann, Halai, Wang, y Craik, 2008) en función de la proporción de resi-

duos de cisteína, su tamaño de bucle y la diversa conectividad de enlaces disulfuro (Kaas y cols., 2012). Dentro de este trabajo se estudiarán y analizarán las superfamilias A, O1 del caracol marino *Conus ermineus*.

Recientes estudios transcriptómicos han demostrado que las superfamilias O1, M, A y T han sido ampliamente muestreadas, mientras que otras superfamilias de genes tienen menos de 30 conotoxinas descritas por cada superfamilia. Esto confirma que las superfamilias O1, M, A y T son en realidad las superfamilias más grandes encontradas hasta la fecha (Akondi y cols., 2014).

La mayor parte de las conotoxinas de la superfamilia A se caracterizan por el marco de cisteína de tipo I (CC-C-C) (Tabla 1) cuyas conectividades de disulfuro son I-III, II-IV, las cuales se dirigen potente y selectivamente a una secuencia de subtipos de nAChR neuronales y neuromusculares. Este conjunto de conotoxinas, que se dirigen a los nAChR, reciben el nombre de  $\alpha$ -conotoxinas, y varias de ellas se han caracterizado en detalle, mostrando su importancia como sondas farmacológicas de la capacidad de los nAChR (Robinson y Norton, 2014).

**Tabla 1**

*Alineación del marco de cisteínas tipo I*

Secuencia
DPCCSNPACNVNNPQIC
QTPGCCWHPACGKNRC
DDCCPDPSCRQNHPELCA

Dentro de esta superfamilia también se encuentran conotoxinas con un marco de cisteína de tipo IV (CC-C-C-C-C) (Tabla 2), las cuales adoptan una conectividad de disulfuro I-V, II-III, IV-VI. Estas conotoxinas se han caracterizado a partir del veneno de los cazadores de peces y

pertenecen a la familia de conopéptidos  $\kappa$ A (Jin y cols., 2019; Robinson y Norton, 2014).

**Tabla 2**

*Alineación del marco de cisteínas tipo IV*

Secuencia
DCCGVKLDMCHPCLCNNSCKQGQGGKRVWEMMKATD
GCCGKYPNAACHPCGCTVGRPPYCDRPSGG
DCCGVKLDMCHPCLCNNSCKQGQGGK

La mayoría de las conotoxinas de la superfamilia O1 se caracterizan por un marco de cisteína tipo VI/VII (Tabla 3), y las que han sido estudiadas hasta el momento adoptan una conectividad de disulfuro I-IV, II-V, III-VI. Hasta la fecha, la superfamilia O1 está compuesta por  $\delta$ -,  $\mu$ -, los cuales son bloqueadores de VGSC;  $\kappa$ - que son bloqueadores de canales de  $K^+$  dependientes de voltaje; y  $\omega$ -conotoxinas corresponden a los bloqueadores de VGCC (Robinson y Norton, 2014).

**Tabla 3**

*Alineación del marco de cisteínas tipo VI/VII*

Secuencia
RPCKPKGRKCFPHQKDCCNKTCTRSKCP
CWSSGTPCGTDSLCCGGCENVSKSKCN
ECGADGAFCLPGLGLNCCSGLCLIVCVST

**Métodos para la reconstrucción de estructura tridimensional de péptidos y proteínas**

La estructura tridimensional de un péptido o conotoxina se define por la disposición de sus residuos de aminoácidos constituyentes en el espacio (Guillén, 2009). Los aminoácidos se unen entre sí a través de enlaces peptídicos, que forman la columna vertebral de la molécula. La actividad biológica y función de las proteínas está determinada por su estructura terciaria y cuaternaria (Guillén, 2009) es por esto que el estudio y los esfuerzos de la comunidad científica

han llevado a la construcción de bases de datos a partir de los resultados obtenidos con el uso de técnicas convencionales como Resonancia Magnética Nuclear (RMN) y Rayos X (RX) de esta forma facilitando la búsqueda y aplicaciones de las mismas (Akondi y cols., 2014). Algunos de los métodos más comunes para la predicción de la estructura tridimensional de proteínas están basados en el uso de campos de fuerza o software basados en predicciones usando inteligencia artificial tales como, AlphaFold2.

### **PEP-FOLD2**

Thevenet y cols. (2012) presentaron una versión actualizada del servidor PEP-FOLD, el cual se basa en un enfoque *de novo* para predecir la estructura tridimensional de péptidos entre 9 y 36 aminoácidos. Este servidor utiliza un alfabeto estructural (por sus siglas en inglés SA) para describir las conformaciones de cuatro residuos consecutivos, para ello predice la serie de letras SA y luego estas son acopladas mediante un procedimiento impulsado por un campo de fuerza de grano grueso (OPEP) (Lamiable y cols., 2016; Maupetit, Derreumaux, y Tuffery, 2009). A partir de la secuencia de aminoácidos, PEP-FOLD realiza una serie de 100 simulaciones y propone la conformación más representativa determinada por energía y población (Maupetit y cols., 2009).

### **AlphaFold2**

Jumper y cols. (2021), en su artículo “Highly accurate protein structure prediction with AlphaFold” publicado en la revista Nature, proponen una versión rediseñada del sistema de inteligencia artificial AlphaFold, siendo este método el que permite predecir la forma de una proteína, a escala y en cuestión de minutos, con precisión atómica, basándose en una recopilación de datos físicos y biológicos sobre la estructura de las proteínas, comparando secuencias de cadenas de

ARN, ADN o estructuras primarias proteicas resaltando sus similitudes, permitiendo encontrar relaciones entre genes o proteínas ya reportados en diferentes bases de datos (Jumper y cols., 2021). AlphaFold2, por su parte promete superar técnicas experimentales como RMN y RX las cuales demandan trabajo y dinero, y superar además otras técnicas computacionales que también utilizan IA, así como otros acercamientos que usan campos de fuerza, y/o métodos cuánticos como por ejemplo Gaussian, Gamess, I-tasser, los cuales participan en el evento de “Evaluación Crítica de la Predicción de la Estructura de Proteínas” (por sus siglas en inglés CASP) (Howes, 2021).

AlphaFold2 es un algoritmo que destaca el poder de los datos de secuencias biológicas y la IA con métodos que permiten predecir la estructura tridimensional de proteínas, así como simular la dinámica molecular y las interacciones entre proteínas, lo cual permite innovar tratamientos personalizados, remodelar y modernizar enfoques para el descubrimiento de fármacos y tratamientos médicos (Nussinov, Zhang, Liu, y Jang, 2022). AlphaFold2 combina numerosos algoritmos basados en “deep learning” para predecir la estructura tridimensional de las proteínas a una resolución de escala experimental o cercana, lo que inspira a la comunidad científica a repensar el estudio de la función, la evolución y la enfermedad (Kuhlman y Bradley, 2019).

Los logros de AlphaFold2 son posibles gracias al Protein Data Bank (PDB), que actualmente cuenta con aproximadamente 200.000 estructuras determinadas experimentalmente. AlphaFold2 ha sido entrenado de acuerdo a las estructuras y secuencias disponible en el PDB, utilizando la secuencia de entrada para consultar las diferentes bases de datos y reconstruir múltiples alineaciones de secuencias (Nussinov y cols., 2022). La predicción tridimensional de la estructura de la proteína se basa: 1). en el modelado basado en plantillas, en el que se utilizan estructuras previamente

determinadas de proteínas relacionadas para modelar la estructura objetivo desconocida; y 2). el modelado sin plantilla, que no se basa en la similitud global con las estructuras en el PDB, sino que simulando interacciones entre aminoácidos y plegamientos ya conocidos propone la posible estructura tridimensional de la secuencia objetivo, por lo cual el modelado sin plantilla se puede aplicar a proteínas o en nuestro caso, péptidos de los que se desconoce su estructura tridimensional y que cuentan con pocas referencias de estructuras tridimensionales similares conocidas (Kuhlman y Bradley, 2019; Nussinov y cols., 2022).

El proceso de predicción de la estructura sin una plantilla suele comenzar con la construcción de un alineamiento de secuencias múltiples de la proteína de interés y secuencias relacionadas. La secuencia diana y sus homólogos se utilizan luego para predecir las características estructurales locales, así como la estructura secundaria y los ángulos de torsión de la columna vertebral, además de características no locales, siendo estas los contactos entre residuos o las distancias entre residuos de la cadena polipeptídica. Estas características predictivas guían el proceso de construcción de un modelo 3D de la estructura de la proteína de interés, que luego se refina, clasifica y compara entre sí para seleccionar la predicción final (Kuhlman y Bradley, 2019).

AlphaFold2 produce una estimación por residuo de su confianza en una escala de 0 a 100. Esta medida de confianza se denomina prueba de diferencia de distancia local predicha (pLDDT), que es una métrica que estima qué tan bien concordaría una estructura predicha con una estructura experimental basada en las coordenadas de carbono de la columna vertebral (Jumper y cols., 2021; Varadi y cols., 2022).

Regiones con pLDDT mayor a 90 se modelan con alta precisión. Las regiones con pLDDT

entre 70 y 90 se modelan con una predicción de columna vertebral generalmente buena. Las regiones con pLDDT entre 50 y 70 son de confianza baja. Las coordenadas tridimensionales de regiones con pLDDT menor a 50 a menudo tienen una apariencia similar a una cinta y no deben interpretarse (Jumper y cols., 2021; Varadi y cols., 2022).

Pese a que en los últimos años se ha acelerado el descubrimiento de secuencias de conotoxinas su caracterización estructural está muy rezagada (Akondi y cols., 2014), lo cual puede deberse a que el uso de las técnicas utilizadas para este fin son demandantes en tiempo y dinero (Jumper y cols., 2021). Es por esto que planteamos como objetivo del presente estudio, evaluar la capacidad de algoritmos computacionales para predecir la estructura tridimensional de péptidos aislados de conotoxinas del caracol *Conus ermineus* al tiempo que se realiza una comparación de las estructuras obtenidas a través de inteligencia artificial (IA), campos de fuerza y datos químicos. De esta manera comprobar nuestra hipótesis de que basándose en el actual conocimiento de que las proteínas son un conjunto de péptidos, el software AlphaFold2, originalmente diseñado para trabajar con proteínas, también puede predecir la estructura tridimensional de péptidos entre 10 y 30 aminoácidos al igual que lo hace para proteínas.

## 1. Objetivos

### 1.1. Objetivo general

Combinar tecnologías emergentes en el campo de la secuenciación y la inteligencia artificial a fin de obtener a partir de secuencias de aminoácidos, estructuras tridimensionales de péptidos con los que se puedan realizar estudios de estructura-actividad.

### 1.2. Objetivos específicos

Determinar si AlphaFold2 puede predecir la estructura tridimensional de un péptido basado en secuencias de 10 a 30 aminoácidos de la misma manera que predice las estructuras tridimensionales de proteínas grandes.

Analizar y comparar las estructuras tridimensionales de los péptidos obtenidos usando software basado en inteligencia artificial (AlphaFold2) y software que no utiliza inteligencia artificial, sino campos de fuerza o *de novo* (PEP-FOLD2).

## 2. Métodos

Se realizó una búsqueda exhaustiva de secuencias de ADN de conotoxinas del caracol *Conus ermineus* a través de la tecnología Illumina (Steeimers y Gunderson, 2005) de secuenciación del archivo europeo de nucleótidos (ENA). Todas las secuencias fueron descargadas y manejadas en formato FASTA. Se escogieron los transcritos correspondientes a las superfamilias A, O1 de conotoxinas las cuales se clasificaron según el número y posición de cisteínas (framework) en familias pertenecientes a cada superfamilia. Seguidamente utilizando el programa ExpASy (Duvaud y cols., 2021) se hizo la traducción de secuencia de transcriptoma a secuencia de aminoácidos dado que es el formato aceptado por los software a emplear.

Una vez obtenido el listado de péptidos a trabajar se realizó una búsqueda informática en las bases de datos tales como Protein Data Bank (PDB) (Burley y cols., 2023), Uniprot (Berman, Henrick, y Nakamura, 2003), Conoserver (Kaas y cols., 2008) y Blast homology para identificar secuencias aminoacídicas similares a las de *Conus ermineus* en dichas superfamilias, para esto se tuvo en cuenta que todas las secuencias presentaran estructura tridimensional reportada ya sea por las técnicas convencionales (RMN, RX) o cualquier otro software y, que presentarán el mismo marco de cisteína y ya sea que pertenecieran o no a esta especie con el fin de tener una estructura referencia para realizar la comparación y saber si los resultados son óptimos.

Empleamos el software AlphaFold2, para modelar la estructura tridimensional de los péptidos, usando las secuencias de aminoácidos elegidas que corresponde a la estructura primaria del

péptido. Con el fin de evaluar la precisión del modelo se realizaron tres réplicas de este proceso, las cuales posteriormente se compararon teniendo en cuenta que todas las réplicas de cada secuencia recuperaran los mismos puentes disulfuro y mantengan la misma estructura tridimensional. Además, se realizó el anterior procedimiento usando el parámetro de campos de fuerza AMBER disponible en el software AlphaFold2.

Para cada entrada, se generaron cinco modelos de estructuras con AlphaFold2 y se seleccionó la estructura mejor clasificada según las puntuaciones de pLDDT, dichas puntuaciones se obtuvieron empleando el software YASARA (Krieger, Vriend, y Spronk, 2013), el cual permite visualizar la estructura coloreada de acuerdo a los valores de B-Factor que corresponden a la puntuación pLDDT de AlphaFold2.

Con el fin de comparar diferentes métodos de reconstrucción de la estructura tridimensional de péptidos y proteínas, reconstruimos las mismas secuencias de aminoácidos en el software PEP-FOLD2, el cual utiliza campos de fuerza y datos químicos (*de novo*) para el modelamiento de la estructura. Una vez hechas todas las réplicas en los software PEP-FOLD2 y AlphaFold2, utilizamos el servidor NOC3.01 para visualizar las estructuras y verificar que haya recuperado de manera correcta todos los enlaces disulfuro.

La comparación de las estructuras tridimensionales recuperadas durante la búsqueda informática en base de datos con las estructuras obtenidas utilizando un software basado en IA y uno basado en campos de fuerzas para la misma secuencia de aminoácidos, se hizo por medio del programa SuperPose Versión 1.022 (Maiti y cols., 2004). Para esta comparación tuvimos en cuenta: que hubiera recuperado el correcto framework de enlaces de disulfuro y el valor de la Raíz de

Desviación Estándar Cuadrada (por sus siglas en inglés RMSD) de las comparaciones, de acuerdo a esto pudimos determinar el software que obtuvo mejores resultados al momento de determinar las estructuras tridimensionales.

Por otra parte empleamos el software AlphaFold2 para determinar la estructura tridimensional del péptido MII, el cual ya contaba con una estructura tridimensional reportada en el PDB obtenida mediante RMN, esto con el fin de comparar la relación entre la identidad y RMSD.

### 3. Resultados

Teniendo en cuenta las secuencias reportadas por Illumina pertenecientes a la especie del caracol marino *Conus ermineus*, se seleccionaron un total de 18 secuencias, 7 de ellas pertenecientes a la superfamilia A y 11 pertenecientes a la superfamilia O1 (Apéndice A).

#### 3.1. AlphaFold2

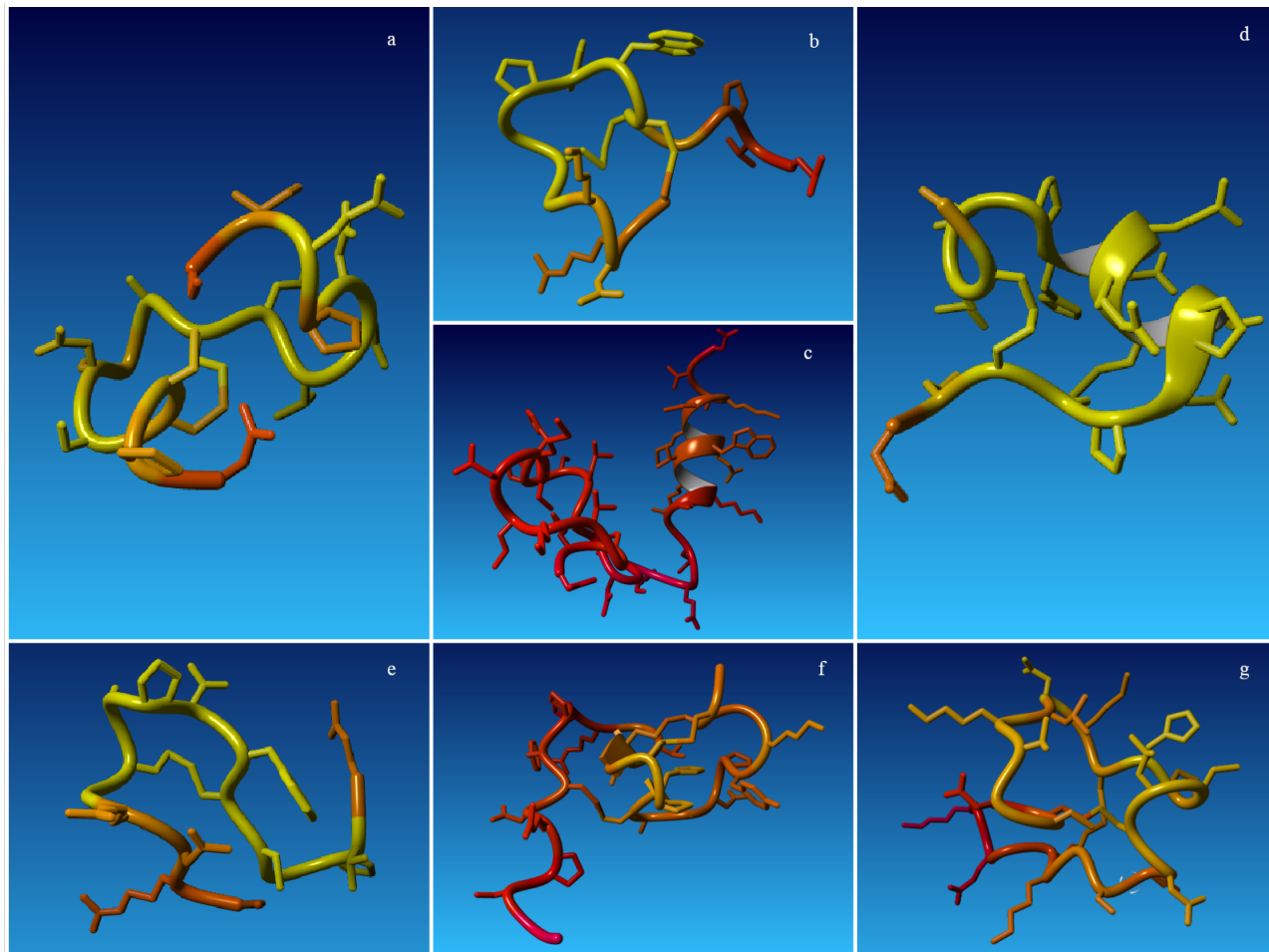
Se encontró que entre las tres réplicas de AlphaFold2 los cinco modelos obtenidos para cada secuencia eran concordantes, recuperando siempre entre ellos los mismos puentes disulfuro. Sin embargo, no siempre la predicción de formación de los puentes disulfuro fue correcta, en algunos casos algunas conotoxinas tenían un par menos de puentes disulfuro, como también formaba los puentes con cisteínas diferentes a lo reportado en teoría o simplemente no formaba ningún enlace de disulfuro. Adicionalmente a esto se valoraron las estructuras recuperadas con el parámetro AMBER, sin embargo, los resultados obtenidos presentaron una menor confianza, por lo tanto se descartaron.

La Figura 1a muestra la estructura tridimensional recuperada mediante AlphaFold2 correspondiente al modelo 3 de la conotoxina A1, sin embargo, a pesar de ser el modelo con mejor valor de pLDDT (Apéndice B) y formar los dos enlaces disulfuro, estos presentaron una formación diferente a la correspondiente al marco de cisteína I (CC-C-C) (Conectividad C1-C3, C2-C4). La Figura 1b muestra la estructura tridimensional recuperada mediante AlphaFold2 que presentó una mejor precisión, correspondiendo al modelo 1 de la conotoxina A2. Este modelo formó todos los

enlaces disulfuro correspondiente al marco de cisteína I (CC-C-C) (Conectividad C1-C3, C2-C4).

### Figura 1

*Estructuras tridimensionales de la superfamilia A obtenidas mediante AlphaFold2*



*Nota.* Regiones coloreadas en amarillo o cercanas a amarillo (Naranjas) presentan una confianza alta, y regiones en rojo o cercanas a rojo (Fucsia) presentan niveles de confianza baja. 1a representa la estructura tridimensional de la conotoxina A1 modelo 3, 1b conotoxina A2 modelo 1, 1c conotoxina A3 modelo 1, 1d conotoxina A4 modelo 1, 1e conotoxina A5 modelo 3, 1f conotoxina A6 modelo 3, 1g conotoxina A7 modelo 3.

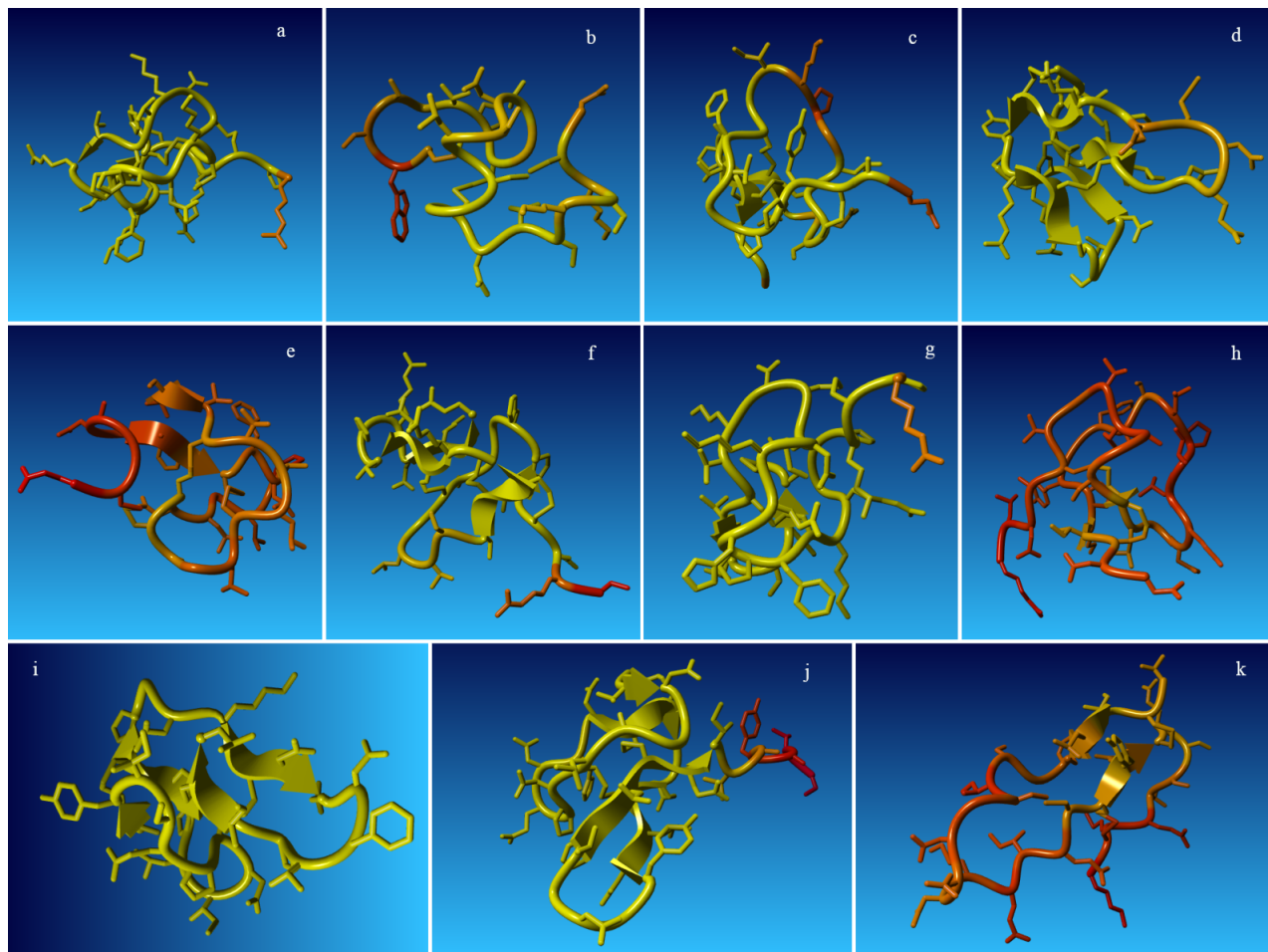
La Figura 1c, 1d, 1e, 1f, 1g, muestran la estructura tridimensional recuperada mediante AlphaFold2 que presentó una mejor precisión, correspondiendo al modelo 1 de la conotoxina A3, el modelo 1 para la A4, el modelo 3 para la A5, el modelo 3 para la A6 y el modelo 3 para la conotoxina A7, respectivamente. Estas predicciones a pesar de presentar el mayor valor de pLDDT (Apéndice B) no formaron todos los puentes disulfuro correspondientes al marco de cisteína.

La Figura 2a, 2c, 2e, 2g, 2i y 2j muestran la estructura tridimensional recuperada mediante AlphaFold2 correspondiente al modelo 3 para la conotoxina O1, el modelo 3 para la conotoxina O3, el modelo 3 de la conotoxina O5, el modelo 3 para la O7, el modelo 4 para la conotoxina O9 y el modelo 5 para la O10, siendo estos los únicos en los que se evidenció la correcta formación de los puentes disulfuros según la teoría, los cuales corresponden al marco de cisteína VI/VII (CC-C-C-C-C) (Conectividad C1-C4, C2-C5, C3-C6).

La Figura 2b, 2d, 2f, 2h y 2k muestran la estructura tridimensional recuperada mediante AlphaFold2 que presentó una mejor precisión, correspondiendo al modelo 2 para la conotoxina O2, el modelo 4 para la O4, el modelo 5 para la conotoxina O6, el modelo 3 para la conotoxina O8, y el modelo 3 para la conotoxina O11, respectivamente. Estas predicciones no formaron todos los puentes disulfuro correspondientes al marco de cisteína VI/VII (CC-C-C-C-C) (Conectividad C1-C4, C2-C5, C3-C6).

**Figura 2**

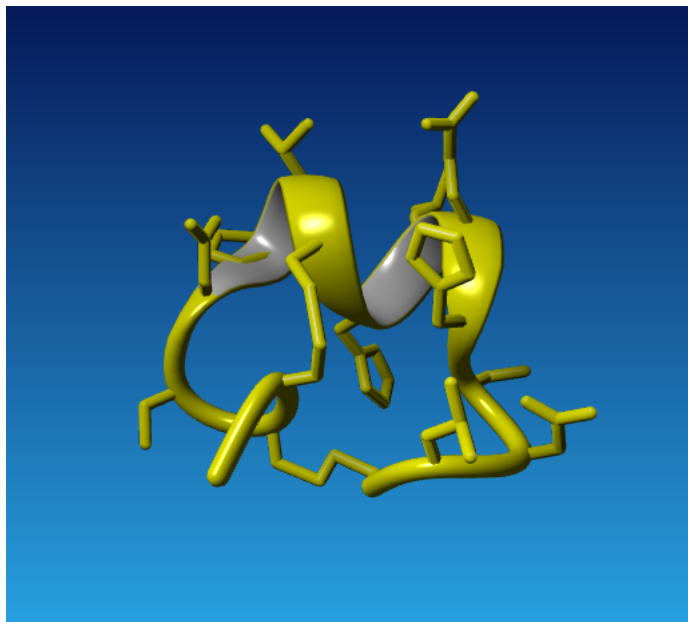
*Estructuras tridimensionales de la superfamilia O obtenidas mediante AlphaFold2*



*Nota.* Regiones coloreadas en amarillo o cercanas a amarillo (Naranjas) presentan una confianza alta, y regiones en rojo o cercanas a rojo (Fucsia) presentan niveles de confianza baja. 2a representa la estructura tridimensional de la conotoxina O1 modelo 3, 2b conotoxina O2 modelo 2, 2c conotoxina O3 modelo 3, 2d conotoxina O4 modelo 4, 2e conotoxina O5 modelo 3, 2f conotoxina O6 modelo 5, 2g conotoxina O7 modelo 3, 2h conotoxina O8 modelo 3, 2i conotoxina O9 modelo 4, 2j conotoxina O10 modelo 5, 2k conotoxina O11 modelo 3.

**Figura 3**

*Estructura tridimensional del péptido Alpha-conotoxin MII obtenida mediante AlphaFold2*



*Nota.* Regiones coloreadas en amarillo o cercanas a amarillo (Naranjas) presentan una confianza alta, y regiones en rojo o cercanas a rojo (Fucsia) presentan niveles de confianza baja.

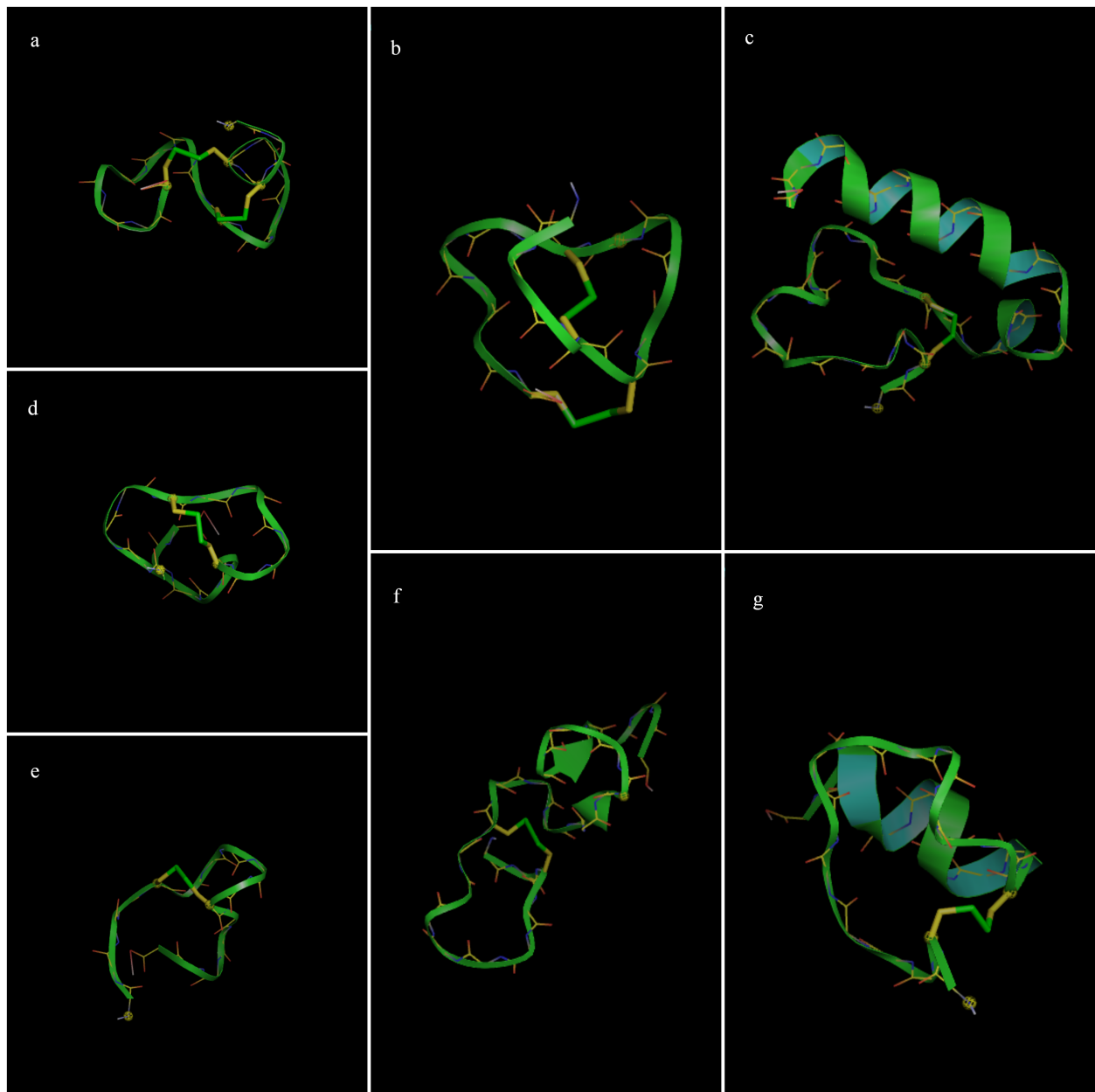
En la Figura 3 se muestra la estructura tridimensional recuperada mediante AlphaFold2 del péptido MII, la cual presentó un valor de pLDDT de 90.8, es decir, presentó una alta precisión.

**3.2. PEP-FOLD2**

Las Figuras 4a, 4b muestran la estructura tridimensional recuperada mediante PEP-FOLD2 correspondiente al modelo 1 de la conotoxina A1 y el modelo 3 para la A2, respectivamente, siendo estos los únicos en los que se evidenció la correcta formación de los puentes disulfuros según la teoría, los cuales corresponden al marco de cisteína I (CC-C-C) (Conectividad C1-C3, C2-C4).

**Figura 4**

*Estructuras tridimensionales de la superfamilia A obtenidas mediante PEP-FOLD2*



*Nota.* 4a representa la estructura tridimensional de la conotoxina A1 modelo 1, 4b conotoxina A2 modelo 3, 4c conotoxina A3 modelo 4, 4d conotoxina A4 modelo 4, 4e conotoxina A5 modelo 2, 4f conotoxina A6 modelo 3, 4g conotoxina A7 modelo 5.

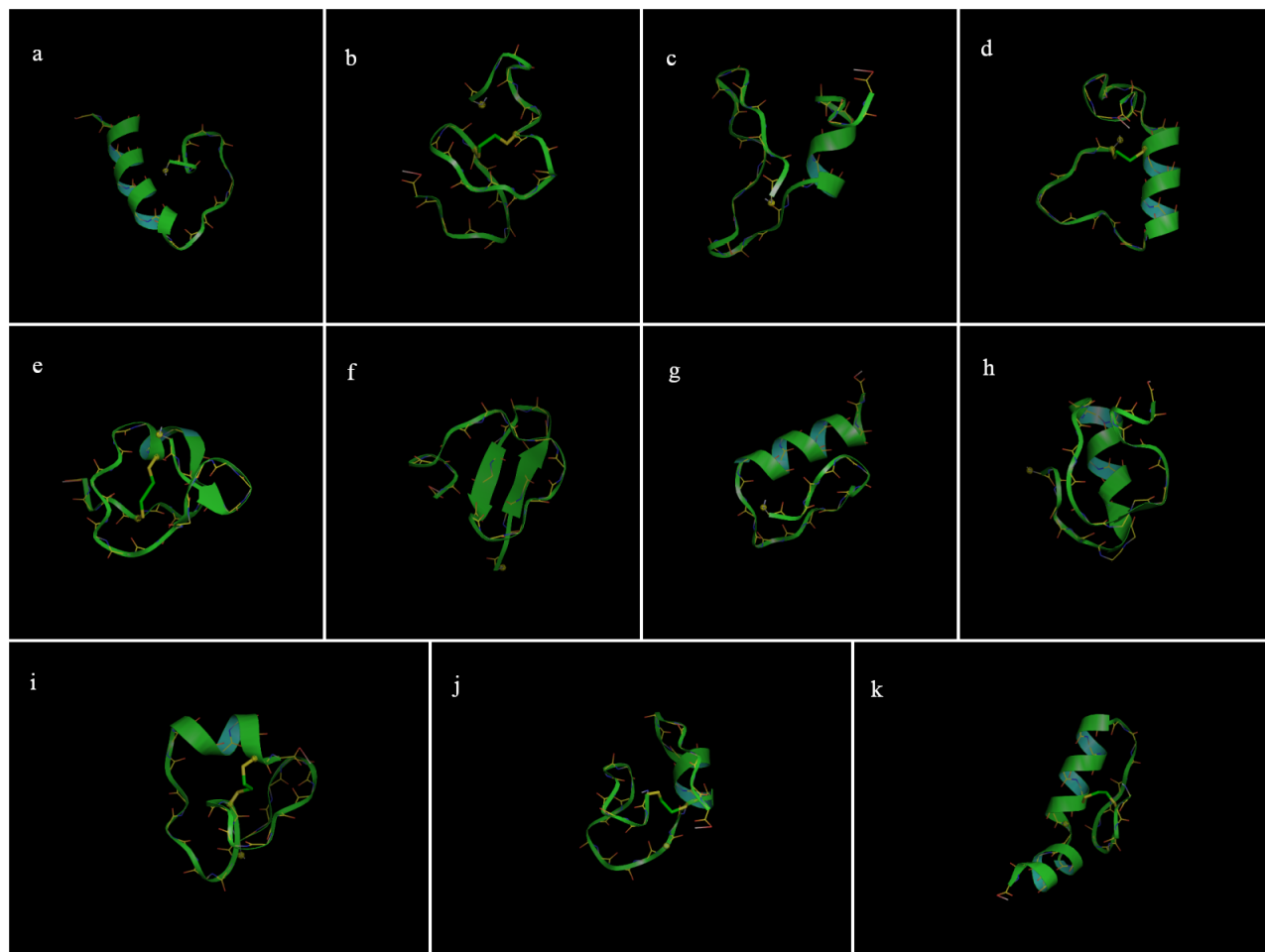
Las Figuras 4c, 4d, 4e, 4f, 4g muestran las estructura tridimensionales recuperadas correspondientes a las conotoxinas A3 modelo 4, A4 modelo 4, A5 modelo 2, A6 modelo 3 y A7 modelo 5, respectivamente. Sin embargo, de los cinco modelos solo estos lograron recuperar un solo enlace de los 3 que según la teoría deben formar al ser conotoxinas pertenecientes al marco de cisteína VI/VII (Conectividad C1-C4, C2-C5, C3-C6).

Las Figuras 5a, 5c, 5f, 5g, 5h muestran las estructuras tridimensionales recuperadas correspondiente a las conotoxinas O1 modelo 3, O3 modelo modelo 2, O6 modelo 1, O7 modelo 5 y O8 modelo 5, respectivamente. Sin embargo, ninguno de los 5 modelos recuperaron el número de puentes disulfuro pertenecientes al marco de cisteína VI/VII (Conectividad C1-C4, C2-C5, C3-C6).

Las figuras 5b, 5d, 5e, 5i, 5j y 5k muestran las estructuras tridimensionales recuperadas correspondientes a las conotoxinas O2 modelo 4, O4 modelo 1, O5 modelo 4, O9 modelo 1, O10 modelo 5 y O11 modelo 4, respectivamente. Sin embargo, de los cinco modelos solo uno de ellos logró recuperar uno de los 3 enlaces pertenecientes al marco de cisteína VI/VII (Conectividad C1-C4, C2-C5, C3-C6).

**Figura 5**

*Estructuras tridimensionales de la superfamilia O obtenidas mediante PEP-FOLD2*



*Nota.* 5a representa la estructura tridimensional de la conotoxina O1 modelo 3, 5b conotoxina O2 modelo 4, 5c conotoxina O3 modelo 2, 5d conotoxina O4 modelo 1, 5e conotoxina O5 modelo 4, 5f conotoxina O6 modelo 1, 5g conotoxina O7 modelo 5, 5h conotoxina O8 modelo 5, 5i conotoxina O9 modelo 1, 5j conotoxina O10 modelo 5, 5k conotoxina O11 modelo 4.

### 3.3. Comparación

Para llevar a cabo la comparación de cada familia de conotoxinas seleccionadas, se eligieron de la base de datos Protein Data Bank conotoxinas aisladas de diferentes familias de caracoles que cumplieran o eran lo más parecidas posibles al marco de cisteínas I y VI/VII (Apéndice C), estas conotoxinas poseían estructura 3D determinada mediante RMN O RX. Se tomaron los modelos recuperados por AlphaFold2 que cumplieran con la correcta formación de las conotoxinas, así estos no correspondieran a la estructura con mayor valor de pDDLT (Apéndice B), dado que este valor de pDDLT se centra en la cadena principal y nuestro interés era la correcta formación de los puentes disulfuro, así mismo, se hizo la comparación con mínimo dos de las estructuras reportadas en el PDB (Apéndice D) y se tuvo en cuenta la que presentó un RMSD total más bajo.

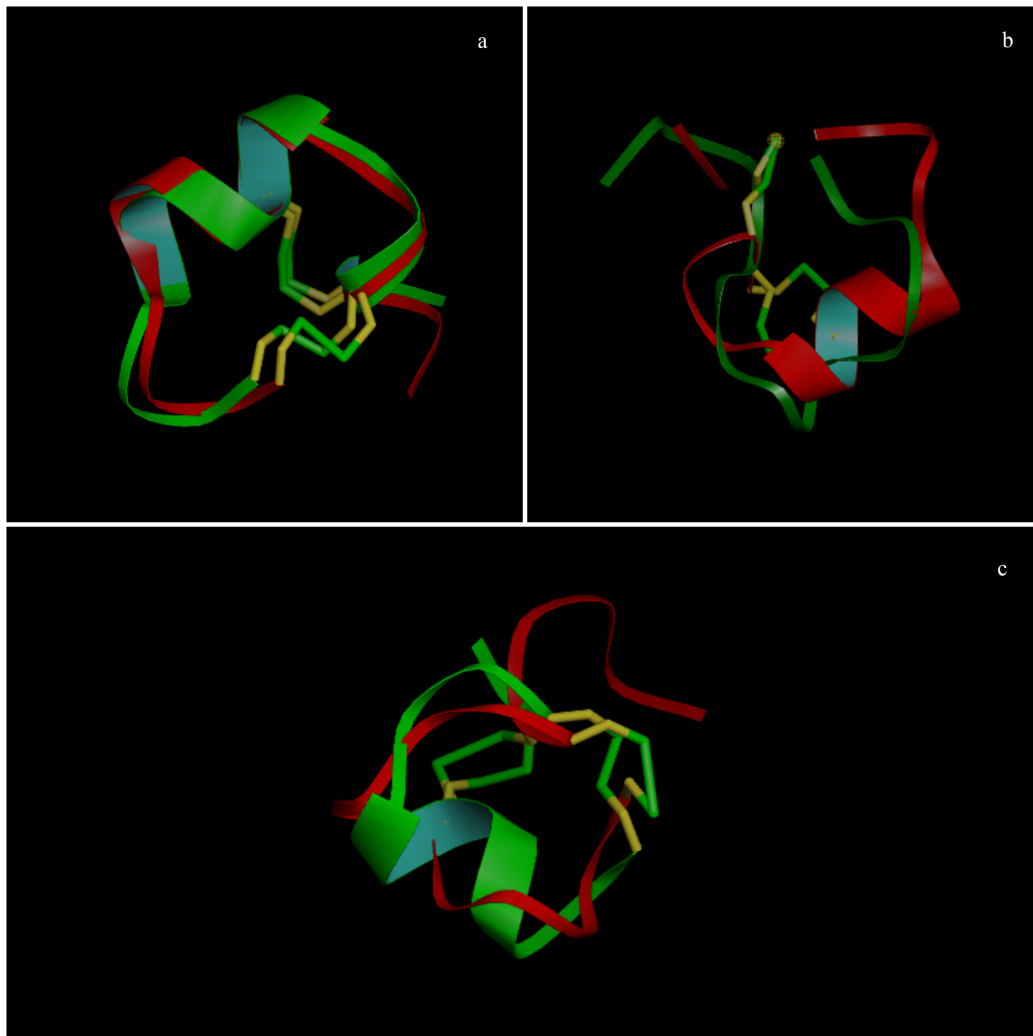
La conotoxina A1 (Figura 6a) mediante el software AlphaFold2 recuperó el modelo 4, el cual fue comparado con las estructuras de los péptidos EI, MII, PIA, PnIB, quienes presentaban valores de identidad distribuidos entre 41.2 y 77.8%, y un RMSD entre 0.95 y 1.96, siendo el péptido PnIB el mejor por su valor más bajo de RMSD. La Conotoxina A2 (Figura 6b) mediante el software AlphaFold2 recuperó los modelos 1 y 3, quienes fueron comparados con las estructuras de los péptidos EI, MII, PIA, y presentaban valores de identidad distribuidos entre 31.6 y 47.4%, y un RMSD entre 2.63 y 2.98, siendo el modelo 3 con el péptido EI el mejor por su RMSD bajo.

La conotoxina A5 (Figura 6c) mediante el software AlphaFold2 recuperó los modelo 1, 2 y 4, los cuales fueron comparados con las estructuras de los péptidos BuIA, BuIA-RX, Vc1a, así mismo, presentaban valores de identidad distribuidos entre 31.6 y 35.7%, y un RMSD entre 2.38 y

3.06, siendo elegido el péptido BuIA-RX para el modelo 2 como el mejor ya que presentó el valor más bajo de RMSD.

**Figura 6**

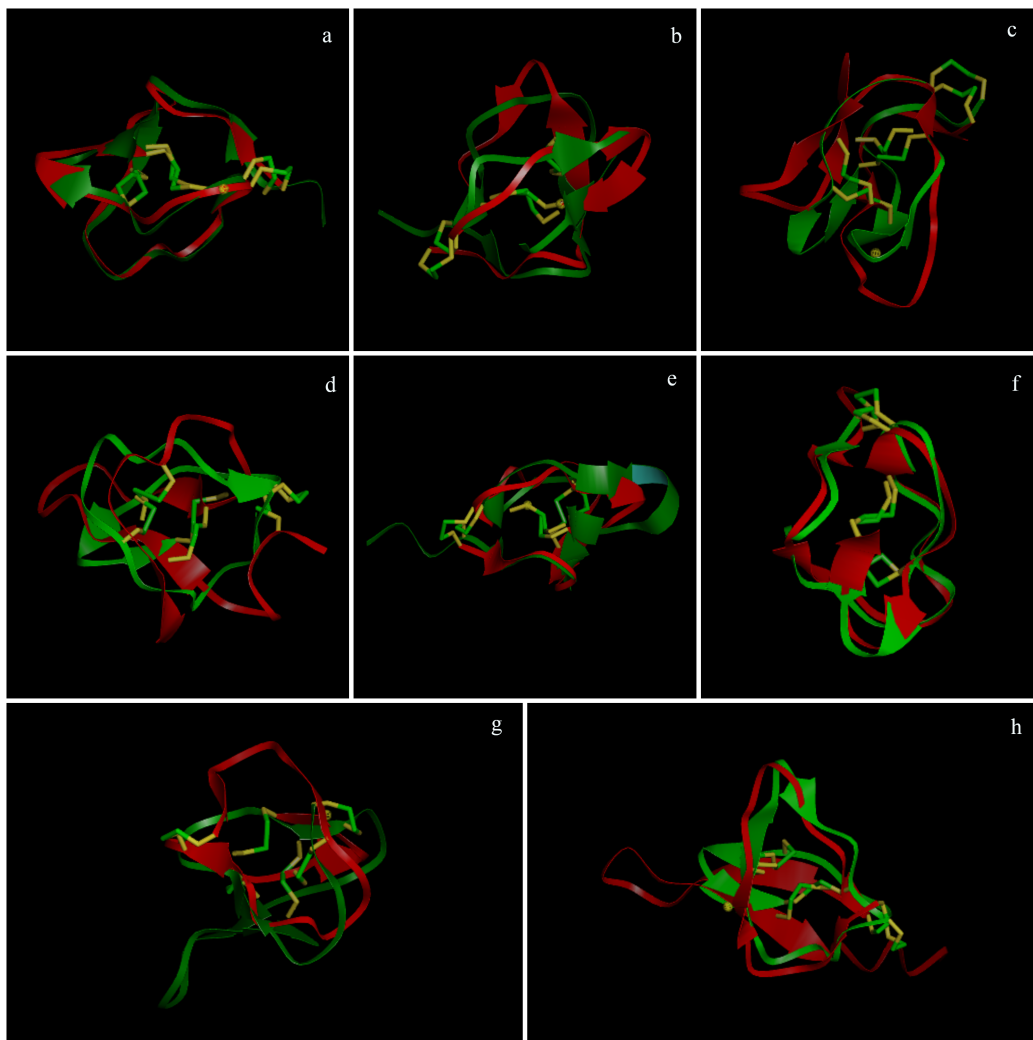
*Estructuras tridimensionales obtenidas de la comparación entre AlphaFold2 y técnicas experimentales (RMN y/o RX) para la superfamilia A*



*Nota.* 6a corresponde a la comparación de entre A1 con PnIB, 6b comparación entre A2 y EI y 6c comparación entre A5 y BuIA.

**Figura 7**

*Estructuras tridimensionales obtenidas de la comparación entre AlphaFold2 y técnicas experimentales (RMN y/o RX) para la superfamilia O*



*Nota.* 7a corresponde a la comparación de entre O1 con MVIIA, 7b comparación entre O3 y BU8, 7c comparación entre O4 y SO-3, 7d comparación entre O5 y MVIIA, 7e comparación entre O6 y MVIIA, 7f comparación entre O7 y MVIIA, 7g comparación entre O9 y SO-3, y 7h comparación entre O10 y SO-3.

La Conotoxina O1 (Figura 7a) mediante el software AlphaFold2 recuperó los modelos 1, 2, 3, 4, 5, quienes fueron comparados con las estructuras de los péptidos BU8, MVIIA, SO-3, presentando valores de identidad distribuidos entre 23.3 y 40.6%, y un RMSD entre 1.55 y 8.84, siendo elegido el péptido MVIIA para el modelo 1 como el mejor ya que presentó el valor más bajo de RMSD.

La Conotoxina O3 (Figura 7b) mediante el software AlphaFold2 recuperó los modelos 1, 2, 3, 4, 5, que fueron comparados con las estructuras de los péptidos BU8, MVIIA, SO-3 presentando valores de identidad distribuidos entre 20 y 40.6%, y un RMSD entre 1.69 y 2.78, siendo elegido el péptido BU8 para el modelo 4 como el mejor ya que presentó el valor más bajo de RMSD.

La conotoxina O4 (Figura 7c) mediante el software AlphaFold2 recuperó los modelo 2, 3, 4 y 5, los cuales fueron comparados con las estructuras de los péptidos MVIIA y SO-3, y presentaron un valor de identidad de 24.2% y valores de RMSD distribuidos entre 6.64 y 6.74 siendo elegido el péptido SO-3 para el modelo 2 como el mejor ya que presentó el valor más bajo de RMSD.

La conotoxina O5 (Figura 7d) mediante el software AlphaFold2 recuperó los modelo 3 y 5, quienes fueron comparados con las estructuras de los péptidos MVIIA y SO-3, presentando un valor de identidad de 32.3% y valores de RMSD distribuidos entre 6.72 y 7.19 siendo elegido el péptido MVIIA para el modelo 3 como el mejor ya que presentó el valor más bajo de RMSD.

La Conotoxina O6 (Figura 7e) mediante el software AlphaFold2 recuperó los modelos 1, 2, 3 los cuales fueron comparados con las estructuras de los péptidos BU8, MVIIA, SO-3 los cuales presentaron valores de identidad distribuidos entre 20 y 23.5%, y un RMSD entre 1.92 y 2.72, siendo elegido el péptido MVIIA para el modelo 2 como el mejor ya que presentaron el valor más

bajo de RMSD.

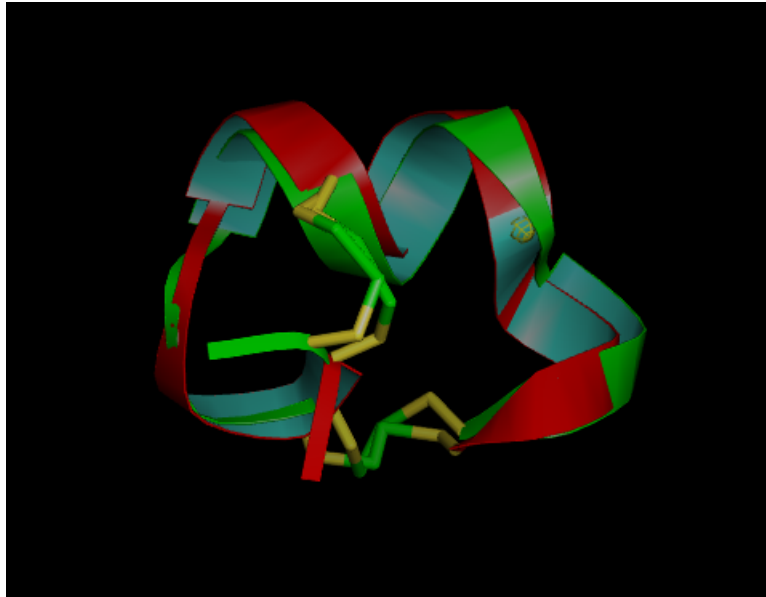
La conotoxina O7 (Figura 7f) mediante el software AlphaFold2 recuperó los modelos 1, 3, 4 y 5, que fueron comparados con las estructuras de los péptidos MVIIA y SO-3, y presentaron valores de identidad distribuidos entre 32.1 y 42.9%, y un RMSD entre 1.55 y 7.51, siendo elegido el péptido MVIIA para el modelo 3 como el mejor ya que presentó el valor más bajo de RMSD.

La Conotoxina O9 (Figura 7g) mediante el software AlphaFold2 recuperó los modelos 1, 2, 3 y 4 los cuales fueron comparados con las estructuras de los péptidos BU8, MVIIA, SO-3 presentando valores de identidad distribuidos entre 32.1 y 42.9%, y un RMSD entre 4.74 y 5.54, siendo elegido el péptido SO-3 para el modelo 1 como el mejor ya que presentó el valor más bajo de RMSD.

La conotoxina O10 (Figura 7h) mediante el software AlphaFold2 recuperó los modelos 1, 2, 3, 4 y 5, quienes fueron comparados con las estructuras de los péptidos MVIIA y SO-3, y presentaron valores de identidad distribuidos entre 16.3 y 20.9%, y un RMSD entre 3.78 y 3.88, siendo elegido el péptido SO-3 para el modelo 5 como el mejor ya que presentó el valor más bajo de RMSD.

**Figura 8**

*Estructura tridimensional de la conotoxina MII obtenida de la comparación entre AlphaFold2 y RMN*



La comparación de la estructura tridimensional de la conotoxina MII (Figura 8) obtenida mediante AlphaFold2 y su respectiva estructura recuperada mediante RMN mostró un porcentaje de identidad y similitud del 100% y un valor de RMSD de 1.72.

## 4. Discusión

### 4.1. AlphaFold2

El estudio de la recuperación de estructuras tridimensionales se ha centrado principalmente en proteínas y no en péptidos, llevando a que no hayan muchas bases de datos disponibles para las estructuras tridimensionales de estos últimos (Akondi y cols., 2014). Por su parte, uno de los métodos que AlphaFold2 utiliza para la recuperación de la estructura tridimensional se basa en el uso de bases de datos (Jumper y cols., 2021), y dada la poca disponibilidad de bases de datos para estructuras de péptidos esto podría afectar la precisión de las estructuras tridimensionales recuperadas para cada superfamilia de conotoxinas.

La predicción de las estructuras pueden basarse en plantillas, es decir, estructuras homólogas a otras ya disponibles en bases de datos, o sin plantillas, es decir, predice la estructura desde cero sin basarse en bases de datos de estructuras homólogas, siendo AlphaFold2 el mejor sistema en la predicción de dominios sin plantillas (Senior y cols., 2020). Sin embargo, Senior y cols. (2020) también aseguran que a pesar de ser el mejor sistema en las predicciones sin plantilla estas rara vez se acercan a la precisión de las estructuras experimentales, y debido a la insuficiencia de bases de datos de estructuras de péptidos, en especial de conotoxinas, se esperaría que AlphaFold2 usara el método sin plantilla para determinar las estructuras estudiadas, lo cual podría estar afectando la precisión de las estructuras recuperadas.

Igualmente, de acuerdo con David, Islam, Tankhilevich, y Sternberg (2022), AlphaFold2

recupera de manera más precisa la cadena principal, adicionalmente, Gulsevin y Meiler (2022), encontraron que AlphaFold2 presenta deficiencias en la predicción de recuperaciones de rotámeros y enlaces disulfuro, lo cual podría estar afectando nuestros resultados ya que las conotoxinas se caracterizan por ser ricas en puentes disulfuro (Jin y cols., 2019).

AlphaFold2 ha sido usado ampliamente para determinar la estructura tridimensional de proteínas, sin embargo, no siempre tienen una alta precisión, como lo reportado por Tunyasuvunakool y cols. (2021), para el proteoma humano donde únicamente recuperó el 58 % de proteínas de forma confiable, adicionalmente, encontraron que en los casos de proteínas que no tienen una estructura totalmente bien definida si no que permanecen sueltas y flexibles, es decir, desordenadas, AlphaFold2 no podría predecir su estructura, ya que de por sí dicha estructura no existe (Ball, 2022). Sin embargo, ellos sugieren que a pesar de que actualmente AlphaFold2 no logra predecir las estructuras desordenadas, en un futuro con suficientes datos y entrenamiento del software, este podría aumentar la precisión en dicha predicción; esto igual podría aplicarse con los péptidos, ya que en el momento su falta de precisión en la recuperación de la estructura podría deberse a como ya se ha mencionado anteriormente, falta de bases de datos, pero si se continúa entrenando el algoritmo para predecir péptidos la precisión aumentaría (Ball, 2022).

Por otra parte, las estructuras tridimensionales de las conotoxinas de la superfamilia A presentaron valores de pLDDT entre un rango de 50 y 70, lo cual de acuerdo con Jumper y cols. (2021); Varadi y cols. (2022), corresponde a una precisión baja. En las estructuras peptídicas correspondientes a la superfamilia O tuvieron una predicción de columna vertebral generalmente buena, ya que los valores de pLDDT oscilaban entre un rango de 70 a 90.

Dentro de la superfamilia A encontramos que las conotoxinas presentaban en su estructura dos y tres puentes disulfuro, encontrándose que las conotoxinas que formaban dos puentes disulfuro (A1, A2, A4, A5) a pesar de tener una precisión baja tuvieron valores de pLDDT más cercanos a 70 que los que presentaban tres puentes disulfuro (A3, A6, A7) cuyos valores de pLDDT fueron más cercanos a 50, lo cual podría concordar con lo propuesto por Gulsevin y Meiler (2022), quienes sugieren que AlphaFold2 tiene deficiencias en la predicción de puentes disulfuro.

Para la superfamilia O todas las conotoxinas presentaban en su estructura tres puentes disulfuro, encontrando que las conotoxinas O5, O8 y O11 presentan una precisión baja e incluso la O8 presenta valores inferiores a 50, lo cual sugiere que su apariencia no está estructurada (Jumper y cols., 2021; Varadi y cols., 2022). Las conotoxinas O1, O2, O3, O4, O6, O7, O9, O10 tuvieron valores de pLDDT entre 70 y 90.

#### **4.2. PEP-FOLD2**

Los resultados mostrados por PEP-FOLD2 no fueron precisos, ya que a pesar de que dentro de los parámetros especificados en el software se definió la correcta formación de los puentes disulfuro, al reconstruir la estructura no recuperó dichos puentes. Por lo tanto, AlphaFold2 como algoritmo de aprendizaje profundo ha demostrado ser más preciso que PEP-FOLD2 que emplea campo de fuerza y datos químicos (*de novo*).

#### **4.3. Comparación**

Como sugiere Carugo y Pongor (2001) los valores de RMSD son un indicador confiable de variabilidad cuando se aplican a proteínas muy similares, es decir, posibles conformaciones de la misma. El valor de RMSD es cero en caso de estructuras idénticas, lo cual casi nunca se

logra en la práctica, así mismo, su valor aumenta conforme disminuya la similaridad entre las estructuras comparadas (Carugo y Pongor, 2001). Lo cual se puede comprobar en la comparación de la estructura tridimensional del péptido MII que al presentar un valor de identidad del 100% su valor de RMSD no fue de cero.

Teniendo en cuenta que el mejor modelo de cada conotoxina se comparó con al menos dos estructuras de RMN o RX que compartían el mismo marco de cisteínas y se presentó la siguiente relación entre identidad y RMSD en el que el valor de identidad fue mayor e igual en mínimo dos de estas comparaciones, sin embargo, difieren en su valor de RMSD, escogiéndose el modelo con menor RMSD. A pesar de tener un mismo valor de identidad, la diferencia en el RMSD radica en que pueden existir conformaciones alternativas debido a la flexibilidad que presentan estos péptidos, es por esto que posiblemente en estos casos la conotoxina adoptará una distribución espacial que coincidiera mejor con una de las estructuras obtenidas mediante RMN o RX, lo cual concuerda con lo dicho por Gulsevin y Meiler (2022), que las estructuras predichas por AlphaFold2 pueden no ser necesariamente incorrectas, sino que simplemente pueden corresponder a una conformación alternativa del péptido, sustentando de esta manera valores tan altos de RMSD, sumando a esto que no se estaban comparando con las mismas secuencias.

Otro factor que pudo afectar los valores de RMSD obtenidos pudo ser la similaridad, ya que pese a conservar las propiedades en los cambios de los aminoácidos como se vio en el ejemplo de la comparación de la conotoxina O1 con respecto a las estructuras de RMN, MVII y SO-3, donde se evidenció para la primera comparación (MVII) un cambio de los aminoácidos Treonina por Serina (Figura 9) y para la segunda comparación (SO-3) tres cambios de aminoácidos, Arginina

por Lisina, Ácido aspártico por Asparagina y Treonina por Serina (Figura 10) a pesar de que comparten la misma polaridad por lo cual se consideran similares estos difieren en el tamaño de sus cadenas laterales, por lo tanto, afecta el valor de RMSD debido a la distribución que pueden tener en el espacio algunos de ellos, es por ello que la primera comparación (MVII) presenta un menor valor de RMSD ya que el único cambio en los aminoácidos no presenta una diferencia muy representativa en su tamaño.

### Figura 9

*Comparación de las conotoxina O1 y MVII*

Conotoxina	Secuencia	Identidad	Similaridad	RMSD
O1	RPCKPKGRKCFPHQKDCCNKTCTRSKCP	23.3%	26.7%	1,55
MVII	CKGKGAKCSR LMYDCCTGSCRSKGC			

### Figura 10

*Comparación de las conotoxina O1 y SO-3*

Conotoxina	Secuencia	Identidad	Similaridad	RMSD
O1	RPCKPKGRKCFPHQKDCCNKTCTRSKCP	23.3%	26.7%	2,15
SO-3	--CKAAGKPCSRIAYNCCTGSCRSKGC-			

Si bien en la mayoría de las comparaciones se cumple la relación RMSD-identidad, en dos de estas, A1 y O3, no se cumple. En la comparación de la conotoxina A1 a pesar de no ser el modelo con mayor valor de identidad presentó un valor de RMSD bajo, lo cual indicaría que la estructura determinada por AlphaFold2 presentó una conformación muy similar a la determinada

mediante RMN. Así mismo, para el caso de la comparación de la estructura del péptido O3, el modelo de mayor identidad no presentó una muy buena precisión respecto a la estructura obtenida con RMN. Sin embargo, son estructuras que no se descartan ya que no se estaban comparando con estructuras idénticas. Cabe resaltar que en estos casos, AlphaFold2 logró formar los puentes disulfuro y en su mayoría predijo las estructuras con alta confianza, lo cual nos lleva a coincidir con Harwitz (2021), quien concuerda con Sameer Velankar en que siguen siendo predicciones que deben confirmarse experimentalmente, sin embargo, no se deben descartar ya que podrían estar aportando información importante relacionada al péptido o proteína y de por sí al mismo organismo, especialmente en casos donde los datos asociados a la estructura son escasos (Ball, 2022; Harwitz, 2021).

De igual manera, con los datos obtenidos concordamos con Ball (2022), quien sugiere que las estructuras ya sean de péptidos o proteínas siguen siendo predicciones, no un hallazgo experimental y a pesar de que su valor de precisión sea alto o bajo, en casos como el de los péptidos, en especial las conotoxinas, donde los datos de sus estructuras tridimensionales son escasos la predicción debe tomarse con confianza, sin embargo, con cautela, ya que puede apoyar futuras investigaciones experimentales, tal como lo expone Ball (2022), junto con Stephen Curry quienes sugieren que las predicciones de AlphaFold2 podrían ser de gran ayuda para la predicción de estructuras muy difíciles de resolver desde cero como es el caso de las predicciones de datos a partir de rayos X.

## 5. Conclusiones

AlphaFold2 recuperó la cadena principal de las estructuras de péptidos entre 10 y 30 aminoácidos con buena precisión, sin embargo, no formó con igual precisión los puentes disulfuro posiblemente debido a que el software no cuenta con suficientes bases de datos relacionadas a péptidos, por lo tanto, se sugiere continuar entrenando el software en la predicción de estructuras tridimensionales de péptidos para de esta manera ampliar su base de datos, apoyado en herramientas como las técnicas experimentales RMN y RX para refinar las estructuras de péptidos generadas por AlphaFold2. Además, encontramos que la relación RMSD-identidad en la mayoría de los casos se cumple, sin embargo, los valores de RMSD fueron bajos pero no iguales a cero, en caso de tener una identidad del 100%. Adicionalmente, comprobamos que dado que nuestra comparación se realizó con estructuras lo más similar posible según lo disponible en bases de datos, en los casos en el que el valor de RMSD fue relativamente bajo podría deberse a que la estructura predicha corresponde a una conformación similar o alternativa del péptido.

Por otra parte, apoyamos la idea de que conocer las predicciones de las estructuras generadas por AlphaFold2 no garantiza que se pueda predecir la relación estructura-actividad ni la afinidad de unión del fármaco pero podría conducir a nuevas hipótesis sobre qué compuesto o fármaco se debe fabricar a continuación (Ball, 2022). Finalmente AlphaFold2 comprueba ser mejor en la predicción de estructuras que el software PEP-FOLD2 que utiliza campo de fuerza o *de novo*.

## **6. Recomendaciones**

AlphaFold2 debe seguirse entrenando en la predicción de estructuras de péptidos logrando enriquecer sus bases de datos y en futuro aumentar la precisión y refinación en estas predicciones. Además de estas dichas predicciones deben seguirse comprobando experimentalmente con otras técnicas aunque presenten un alto valor de precisión.

### Referencias

- Akondi, K. B., Muttenthaler, M., Dutertre, S., Kaas, Q., Craik, D. J., Lewis, R. J., y Alewood, P. F. (2014). Discovery, synthesis, and structure–activity relationships of conotoxins. *Chemical Reviews*, *114*(11), 5815–5847.
- Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., De Castro, E., ... Heinz, S. (2012). Expasy: Sib bioinformatics resource portal. *Nucleic acids research*, *40*(W1), W597–W603.
- Ball, P. (2022). Deepmind has predicted the shape of every protein known to science. how excited should we be? *Prospect Magazine*, *Obtenido de <http://https://www.prospectmagazine.co.uk/ideas/technology/38896/deepmind-has-predicted-the-shape-of-every-protein-known-to-science.-how-excited-should-we-be>*.
- Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., ... Humphray, A. J., Sean J ... Smith (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *nature*, *456*(7218), 53–59.
- Berman, H., Henrick, K., y Nakamura, H. (2003). Announcing the worldwide protein data bank. *Nature Structural & Molecular Biology*, *10*(12), 980–980.
- Bodanszky, M. (1988). Peptide chemistry. *A Practical Textbook*,.
- Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chao, H., Chen, L., ... Guranovic, C., Vladimir ... Zardecki (2023). Rcsb protein data bank (rcsb.org): delivery of experimentally-

- determined pdb structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Research*, 51(D1), D488–D508.
- Cang, H., Chen, M., y Nymeyer, H. (2006). Noc: a free molecular explorer for protein structure visualization, validation and analysis. *Florida: Nymeyer's Group, Institute of Molecular Biology, Florida State University*.
- Carugo, O., y Pongor, S. (2001). A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein science*, 10(7), 1470–1473.
- Carvajal Patiño, D. F. (2015). Un método híbrido para la predicción de la estructura terciaria de las proteínas a partir de su secuencia de aminoácidos. *Ingeniería de Sistemas, Universidad Nacional de Colombia*.
- Chen, J., Liu, X., Yu, S., Liu, J., Chen, R., Zhang, Y., ... Dai, Q. (2021). A novel  $\omega$ -conotoxin bu8 inhibiting n-type voltage-gated calcium channels displays potent analgesic activity. *Acta Pharmaceutica Sinica B*, 11(9), 2685–2693.
- Chi, S.-W., Kim, D.-H., Olivera, B. M., McIntosh, J. M., y Han, K.-H. (2006). Nmr structure determination of  $\alpha$ -conotoxin buia, a novel neuronal nicotinic acetylcholine receptor antagonist with an unusual 4/4 disulfide scaffold. *Biochemical and biophysical research communications*, 349(4), 1228–1234.
- Chi, S.-W., Lee, S.-H., Kim, D.-H., Kim, J.-S., Olivera, B. M., McIntosh, J. M., y Han, K.-H. (2005). Solution structure of  $\alpha$ -conotoxin pia, a novel antagonist of  $\alpha 6$  subunit containing nicotinic acetylcholine receptors. *Biochemical and biophysical research communications*, 338(4), 1990–1997.

- Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., ... Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read smrt sequencing data. *Nature methods*, 10(6), 563–569.
- Clark, R. J., Fischer, H., Nevin, S. T., Adams, D. J., y Craik, D. J. (2006). The synthesis, structural characterization, and receptor specificity of the  $\alpha$ -conotoxin vc1. 1. *Journal of Biological Chemistry*, 281(32), 23254–23263.
- Dao, F.-Y., Yang, H., Su, Z.-D., Yang, W., Wu, Y., Ding, H., ... Lin, H. (2017). Recent advances in conotoxin classification by using machine learning methods. *Molecules*, 22(7), 1057.
- David, A., Islam, S., Tankhilevich, E., y Sternberg, M. J. (2022). The alphafold database of protein structures: a biologist's guide. *Journal of molecular biology*, 434(2), 167336.
- Duvaud, S., Gabella, C., Lisacek, F., Stockinger, H., Ioannidis, V., y Durinx, C. (2021). Expasy, the swiss bioinformatics resource portal, as designed by its users. *Nucleic Acids Research*, 49(W1), W216–W227.
- Guillén, M. V. L. (2009). Estructura y propiedades de las proteínas. *Obtenido de [http://www. uv. es](http://www.uv.es): [http://www. uv. es/tunon/pdf\\_doc/proteinas\\_09. pdf](http://www. uv. es/tunon/pdf_doc/proteinas_09. pdf). 34p.*
- Gulsevin, A., y Meiler, J. (2022). Benchmarking peptide structure prediction with alphafold2. *bioRxiv*, 2022–02.
- Harwitz, E. (2021). *Alphafold 'pushes science forward' by releasing structures of almost all human proteins* (Vol. 99) (n.º 28).
- Hill, J. M., Oomen, C. J., Miranda, L. P., Bingham, J.-P., Alewood, P. F., y Craik, D. J. (1998). Three-dimensional solution structure of  $\alpha$ -conotoxin mii by nmr spectroscopy: effects of

- solution environment on helicity. *Biochemistry*, *37*(45), 15621–15630.
- Howes, L. (2021). *Protein structure prediction reached a new level* (Vol. 117) (n.º 12). Amer Inst Chemical Engineers 3 Park Ave, New York, NY 10016-5901 USA.
- Hu, S.-H., Gehrmann, J., Alewood, P. F., Craik, D. J., y Martin, J. L. (1997). Crystal structure at 1.1 Å resolution of  $\alpha$ -conotoxin pnib: comparison with  $\alpha$ -conotoxins pnia and gi. *Biochemistry*, *36*(38), 11323–11330.
- Jin, A.-H., Muttenthaler, M., Dutertre, S., Himaya, S., Kaas, Q., Craik, D. J., ... Alewood, P. F. (2019). Conotoxins: Chemistry and biology. *Chemical reviews*, *119*(21), 11510–11549.
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... Hassabis, D. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, *596*(7873), 583–589.
- Kaas, Q., Westermann, J.-C., Halai, R., Wang, C. K., y Craik, D. J. (2008). Conoserver, a database for conopeptide sequences and structures. *Bioinformatics*, *24*(3), 445–446.
- Kaas, Q., Yu, R., Jin, A.-H., Dutertre, S., y Craik, D. J. (2012). Conoserver: updated content, knowledge, and discovery tools in the conopeptide database. *Nucleic acids research*, *40*(D1), D325–D330.
- Kohno, T., Kim, J. I., Kobayashi, K., Kodera, Y., Maeda, T., y Sato, K. (1995). Three-dimensional structure in solution of the calcium channel blocker. omega.-conotoxin mviiia. *Biochemistry*, *34*(32), 10256–10265.
- Krieger, E., y Vriend, G. (2014). Yasara view—molecular graphics for all devices—from smartphones to workstations. *Bioinformatics*, *30*(20), 2981–2982.

- Krieger, E., Vriend, G., y Spronk, C. (2013). Yasara—yet another scientific artificial reality application. *YASARA org*, 993, 51–78.
- Kuhlman, B., y Bradley, P. (2019). Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology*, 20(11), 681–697.
- Kukurba, K. R., y Montgomery, S. B. (2015). Rna sequencing and analysis. *Cold Spring Harbor Protocols*, 2015(11), pdb-top084970.
- Lamiable, A., Thévenet, P., Rey, J., Vavrusa, M., Derreumaux, P., y Tufféry, P. (2016). Pep-fold3: faster de novo structure prediction for linear peptides in solution and in complex. *Nucleic acids research*, 44(W1), W449–W454.
- Maiti, R., Van Domselaar, G. H., Zhang, H., y Wishart, D. S. (2004). Superpose: a simple server for sophisticated structural superposition. *Nucleic acids research*, 32(suppl\_2), W590–W594.
- Maupetit, J., Derreumaux, P., y Tuffery, P. (2009). Pep-fold: an online resource for de novo peptide structure prediction. *Nucleic acids research*, 37(suppl\_2), W498–W503.
- Nussinov, R., Zhang, M., Liu, Y., y Jang, H. (2022). Alphafold, artificial intelligence (ai), and allostery. *The Journal of Physical Chemistry B*, 126(34), 6372–6383.
- Park, K.-H., Suk, J.-E., Jacobsen, R., Gray, W. R., McIntosh, J. M., y Han, K.-H. (2001). Solution conformation of  $\alpha$ -conotoxin ei, a neuromuscular toxin specific for the  $\alpha 1/\delta$  subunit interface of torpedo nicotinic acetylcholine receptor. *Journal of Biological Chemistry*, 276(52), 49028–49033.
- Robinson, S. D., y Norton, R. S. (2014). Conotoxin gene superfamilies. *Marine drugs*, 12(12), 6058–6101.

- Rudiño-Piñera, E., Quintero-Hernández, V., y Juárez-González, V. R. (2022). El protein data bank (pdb) y su impacto en la investigación científica.
- Salomon-Ferrer, R., Case, D., y Walker, R. (2013). *An overview of the amber biomolecular simulation package. wires comput mol sci* 3: 198–210.
- Sargsyan, K., Grauffel, C., y Lim, C. (2017). How molecular size impacts rmsd applications in molecular dynamics simulations. *Journal of chemical theory and computation*, 13(4), 1518–1524.
- Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., ... Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792), 706–710.
- Shen, W., Le, S., Li, Y., y Hu, F. (2016). Seqkit: a cross-platform and ultrafast toolkit for fasta/q file manipulation. *PloS one*, 11(10), e0163962.
- Stemmers, F. J., y Gunderson, K. L. (2005). Illumina, inc. *Pharmacogenomics*, 6(7), 777–82.
- Thevenet, P., Shen, Y., Maupetit, J., Guyon, F., Derreumaux, P., y Tuffery, P. (2012). Pep-fold: an updated de novo structure prediction server for both linear and disulfide bonded cyclic peptides. *Nucleic acids research*, 40(W1), W288–W293.
- Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., ... Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873), 590–596.
- Udayangani, S. (2019). Difference between similarity and identity in sequence alignment. <https://www.differencebetween.com/difference-between-similarity-and-identity->

*in-sequence-alignment/*: :text=Hence%2C%20similarity%20and%20identity%20are, exactly%20between%20two%20different%20sequences.

- Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., ... Velankar, S. (2022). Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1), D439–D444.
- Xu, Y., Xu, D., Liang, J., y otros. (2007). *Computational methods for protein structure prediction and modeling* (Vol. 2). Springer.
- Yan, Y., Tu, G., Luo, X., Dai, Q., Huang, P., y Zhang, R. (2003). Three-dimensional solution structure of  $\omega$ -conotoxin so3 determined by 1 h nmr. *Chinese Science Bulletin*, 48, 1097–1102.