9. INGENIERÍA FUNCIONAL Y APRENDIZAJE PROFUNDO EN BIOINFORMÁTICA PARA LA PREDICCIÓN DE ESTRUCTURAS DE PROTEÍNAS[51]

Functional Engineering and Deep Learning in Bioinformatics for the Prediction of Protein Structures

Jordan Piero Borda Colque[52]

Fred Torres-Cruz[53]

Leonel Coyla Idme[54]

Juan Kenyhy Hancco Quispe[55]

Hugo Ticona Salluca[56]

Pares evaluadores: Red de Investigación en Educación, Empresa y Sociedad – REDIEES.[57]

[51] Derivado del proyecto de investigación: Ingeniería Funcional y Aprendizaje Profundo en Bioinformática para la predicción de estructuras de proteínas.

[52] Instituto de Investigación en Inteligencia Computacional y Ciencia de Datos, Departamento Académico de Ingeniería Estadística e Informática, Universidad Nacional del Altiplano de Puno, P.O. Box 291, Puno-Perú, https://orcid.org/0000-0001-8488-1658 , jordanpieroborda@gmail.com

[53] Instituto de Investigación en Inteligencia Computacional y Ciencia de Datos, Departamento Académico de Ingeniería Estadística e Informática, Universidad Nacional del Altiplano de Puno, P.O. Box 291, Puno-Perú, https://orcid.org/0000-0003-0834-6834, ftorres@unap.edu.pe

[54] Instituto de Investigación en Inteligencia Computacional y Ciencia de Datos, Departamento Académico de Ingeniería Estadística e Informática, Universidad Nacional del Altiplano de Puno, P.O. Box 291, Puno-Perú, https://orcid.org/0000-0003-3538-1061 , lcoyla@unap.edu.pe

[55] Instituto de Investigación en Inteligencia Computacional y Ciencia de Datos, Departamento Académico de Ingeniería Estadística e Informática, Universidad Nacional del Altiplano de Puno, P.O. Box 291, Puno-Perú, https://orcid.org/0000-0002-2125-0530 , jkenyhyhq@gmail.com

[56] Instituto de Investigación en Inteligencia Computacional y Ciencia de Datos, Departamento Académico de Ingeniería Estadística e Informática, Universidad Nacional del Altiplano de Puno, P.O. Box 291, Puno-Perú, https://orcid.org/0000-0002-3800-8433 , hts.ez.v@gmail.com

[57] Red de Investigación en Educación, Empresa y Sociedad – REDIEES. www.rediees.org

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

INGENIERÍA FUNCIONAL Y APRENDIZAJE PROFUNDO EN BIOINFORMÁTICA PARA LA PREDICCIÓN DE ESTRUCTURAS DE PROTEÍNAS

Jordan Piero Borda Colque, Fred Torres-Cruz, Leonel Coyla Idme, Juan Kenyhy Hancco Quispe, Hugo Ticona Salluca

RESUMEN

La predicción de la estructura terciaria de proteínas es un desafío altamente complejo y de larga data en el campo de la bioinformática estructural. Los métodos tradicionales de predicción de estructuras se basan en la descomposición del proceso en múltiples subproblemas más manejables, que incluyen la predicción de estructuras locales, la estimación de mapas de contactos, el ensamblaje de fragmentos, el refinamiento y la evaluación de la calidad. En los últimos años, se han incorporado técnicas de aprendizaje profundo para mejorar la precisión en la predicción de estructuras de proteínas. Estas técnicas han demostrado ser particularmente exitosas en la resolución de subproblemas como la estimación de mapas de contactos y estructuras secundarias. La calidad de los mapas de contacto predichos, así como sus variantes, como los mapas de distancia y de orientación, ha mejorado significativamente el rendimiento en la predicción de estructuras terciarias. En la última competencia CASP, se presentaron modelos de redes neuronales profundas de extremo a extremo que mejoraron significativamente la calidad de las estructuras predichas, lo que indica que las técnicas de aprendizaje profundo son una herramienta prometedora para abordar este problema complejo. En este capítulo, se describe el progreso reciente en el desarrollo y aplicaciones de técnicas de aprendizaje profundo para la predicción de estructuras de proteínas, y se discuten posibles razones de su efectividad.

Palabras Clave: bioinformática; aprendizaje profundo; mapas de contacto; estructuras secundarias; estructuras terciarias; redes neuronales profundas.

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

ABSTRACT

Tertiary structure prediction of proteins is an arduous and long-standing challenge that has posed a major obstacle in the field of structural bioinformatics for many years. The conventional methods for predicting protein structures tend to break down the process into multiple, more manageable subproblems, such as the prediction of local structures, the estimation of contact maps, the assembly of fragments, the refinement of predicted structures, and the assessment of quality. In recent years, deep learning techniques have been integrated into the protein structure prediction pipeline to enhance its accuracy, and have proven to be particularly effective in resolving subproblems related to the estimation of contact maps and secondary structures. The quality of predicted contact maps, along with their various forms such as distance and orientation maps, has greatly improved the performance of the tertiary structure prediction process. Notably, end-to-end deep neural network models presented at the most recent CASP competition have made significant strides in enhancing the quality of predicted structures, highlighting the tremendous potential of deep learning techniques as a powerful tool to tackle this complex problem. This chapter aims to present the recent progress in the development and application of deep learning techniques for protein structure prediction, and to examine potential reasons for their remarkable effectiveness.

Keywords: bioinformatics, deep learning, contact maps, secondary structures, tertiary structures; deep neural networks.

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

INTRODUCCIÓN

Las proteínas son macromoléculas biológicas cruciales, compuestas por una o varias cadenas polipeptídicas que se pliegan en diversas estructuras tridimensionales en el espacio. Dado que las estructuras de las proteínas tienen un gran impacto en sus funciones y en sus interacciones con otras moléculas, se han diseñado y empleado diversos métodos experimentales para resolver sus estructuras, tales como la cristalografía de rayos X, la espectroscopia de resonancia magnética nuclear (RMN) y la microscopía crioelectrónica (crio-EM). No obstante, estos métodos suelen resultar costosos, lentos y laboriosos.

La determinación computacional de la estructura de la proteína, también conocida como predicción de la estructura terciaria de proteínas, proporciona información de bajo costo que complementa y respalda estos esfuerzos experimentales. El objetivo principal de la predicción de la estructura terciaria de proteínas es derivar las coordenadas tridimensionales de cada átomo pesado de una proteína dada a partir de su secuencia de aminoácidos. Su fundamento teórico se basa en la hipótesis termodinámica de que la estructura nativa de una proteína se puede determinar por su secuencia de aminoácidos en el entorno fisiológico estándar (Anfinsen, 1973). Aunque algunas proteínas muestran cambios en su plegamiento (Porter, 2018; Pauwels, 2007) y algunas se consideran intrínsecamente desordenadas (Oldfield, 2019; Lietaud, 2016), se podría afirmar que la mayoría de las proteínas siguen el paradigma anterior.

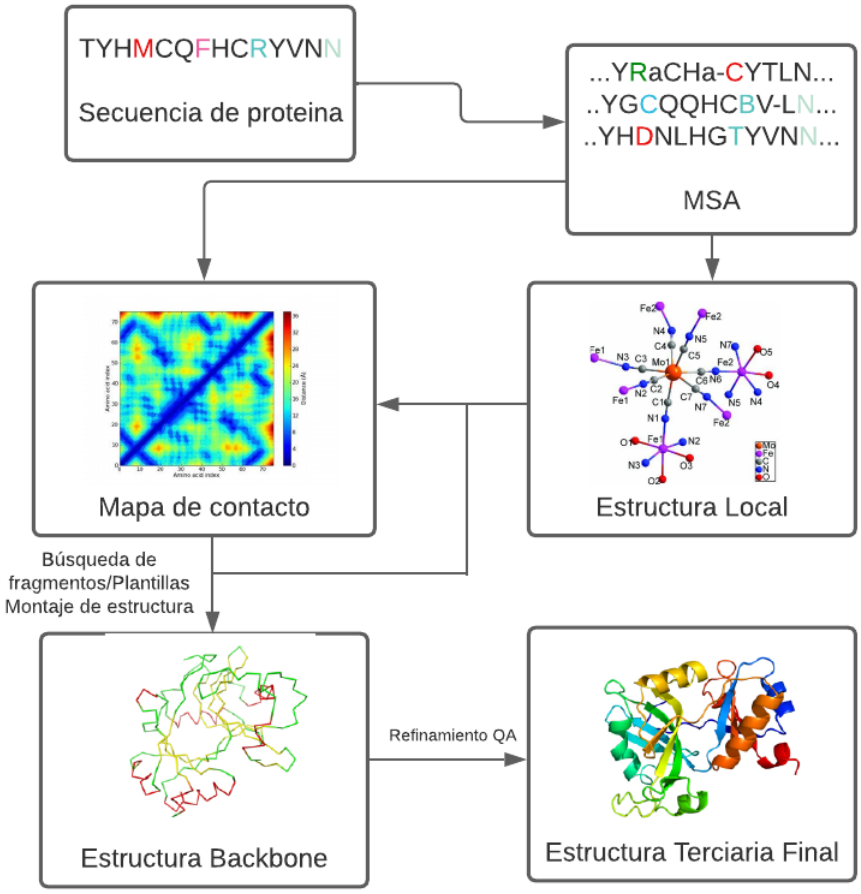
Durante décadas, se han realizado múltiples esfuerzos para predecir la estructura de las proteínas. La mayoría de los enfoques se pueden clasificar en dos tipos: el modelado basado en plantillas (TBM, por sus siglas en inglés) y el modelado libre (FM, por sus siglas en inglés) (Zhang, 2008). El TBM requiere de la proteína consultada que tenga proteínas homólogas con estructuras conocidas. El FM, también conocido como modelado ab initio, no requiere plantillas con estructuras muy similares. Conforme se proponen más algoritmos híbridos que combinan los dos tipos de enfoques de modelado, el límite entre TBM y FM se vuelve borroso (Yang, 2015; Pearce, 2021). El procedimiento adoptado por muchos enfoques para la predicción de la estructura de la proteína generalmente consta de varios módulos,

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

como la predicción de la estructura local, la predicción del mapa de contactos, el ensamblaje de fragmentos, el refinamiento y la evaluación de la calidad (ver Figura 1).

Figura 1

Procedimientos previos a la aplicación de redes neuronales



Nota: elaboración propia.

La figura 1 ilustra los procedimientos previos a la aplicación de redes neuronales profundas de extremo a extremo para la predicción de estructuras de proteínas, los cuales son considerados relativamente independientes. Para una secuencia de proteína específica, se emplean herramientas de alineación de secuencias para obtener su alineamiento múltiple de secuencias (MSA) a través de la búsqueda en bases de datos de secuencias. Posteriormente, se realizan predicciones de estructuras locales y mapas de contactos basados en el MSA, los

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

cuales pueden ser utilizados para ensamblar la estructura de la red troncal. El proceso de refinamiento es empleado para añadir las cadenas laterales y eliminar cualquier violación estructural y choques presentes. Finalmente, se aplica la evaluación de calidad (QA) para estimar la precisión de los señuelos y determinar la estructura terciaria final.

Las estructuras de las proteínas son una descripción geométrica de un fragmento continuo de polipéptido, que incluye las estructuras secundarias. Los mapas de contacto son una medida de la distancia euclidiana entre residuos en una proteína. Si se predice con precisión la estructura local y el mapa de contacto, la búsqueda de conformaciones correspondientes se reduce significativamente. Los fragmentos pueden ensamblarse mediante algoritmos heurísticos que siguen principios físicos o reglas empíricas, lo cual se considera un enfoque común. Sin embargo, las técnicas de aprendizaje profundo de extremo a extremo han introducido un nuevo enfoque. El aprendizaje profundo es un subcampo del aprendizaje automático y del aprendizaje de representación que extrae características de alto nivel y realiza predicciones simultáneamente (Rengio, 2013). Los modelos superficiales tradicionales dependen en gran medida de la ingeniería de características, mientras que los algoritmos de aprendizaje profundo se consideran modelos de extremo a extremo. El núcleo de las técnicas de aprendizaje profundo es la red neuronal artificial (ANN), que se inspira biológicamente en el cerebro humano (Alom, 2018). Desde el punto de vista del aprendizaje automático, la ANN puede considerarse como un perceptrón multicapa (MLP) con función de activación no lineal y retro propagación de errores. Las características aprendidas mediante transformaciones no lineales múltiples son más discriminantes y compactas en comparación con las características manuales. Desde que AlexNet ganó el ImageNet Challenge en 2012 (Krizhevsky, 2012), el aprendizaje profundo ha mostrado un rendimiento muy prometedor en el reconocimiento de voz, el reconocimiento de imágenes y el procesamiento del lenguaje natural (NLP) (Amodei, 2016). Con el desarrollo de técnicas de secuenciación de alto rendimiento, se acumulan rápidamente secuencias genéticas y de proteínas. El mayor conjunto de datos de secuencias de proteínas, BFD, contiene alrededor de 2500 millones de secuencias derivadas de metagenomas. En los últimos años, los investigadores han aplicado técnicas de aprendizaje profundo a la biología molecular, especialmente a la biología estructural. El éxito de RaptorX-Contact en la competencia Critical Assessment of protein Structure Prediction (CASP) mostró que el aprendizaje

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

profundo se desempeña mejor en la predicción del mapa de contactos de proteínas cuando se dispone de datos de secuencias de proteínas a gran escala (Alipanahi, 2015). Posteriormente, se propusieron métodos más profundos basados en el aprendizaje para una variedad de estructuras de proteínas basadas en secuencias. En comparación con los métodos tradicionales de aprendizaje automático, la mayor precisión de predicción lograda por los métodos basados en el aprendizaje profundo puede deberse a su poderosa capacidad de representación. En CASP13, AlphaFold, que aplicó redes neuronales profundas al mapa de contactos y la predicción del ángulo de torsión, mostró resultados prometedores. La red geométrica recurrente (RGN) de extremo a extremo también se propuso en CASP13 (AlQuraishi, 2019).

Los modelos de lenguaje de proteínas se utilizan para aprender información evolutiva a partir de tareas de predicción no supervisadas del tipo de aminoácidos (Rives, 2019). Las incrustaciones de secuencias de proteínas aprendidas se han demostrado más efectivas para muchas tareas de predicción de estructuras posteriores que las características derivadas directamente del perfilado de secuencias o del análisis de acoplamiento directo (DCA) (Weigt, 2009). En la última evaluación de predicción de estructuras CASP 14, se presentó un modelo integral, denominado AlphaFold2, que utiliza datos etiquetados y no etiquetados mediante aprendizaje autosupervisado y aumento de datos (Jumper, 2021). La mejora aportada por AlphaFold2 es significativa y demuestra que el aprendizaje profundo es una herramienta prometedora para la predicción de la estructura de proteínas en el futuro. En este capítulo, se revisan los avances recientes en modelos de aprendizaje profundo para la predicción de estructuras de proteínas basados en secuencias. En primer lugar, se presentan varias arquitecturas de redes neuronales profundas que se han utilizado ampliamente en la bioinformática estructural. A continuación, se discute el último modelo de lenguaje de proteínas y sus efectos en tareas de predicción de la estructura de proteínas. Posteriormente, se comparan y analizan en detalle los modelos de predicción de estructuras secundarias supervisados existentes, los modelos de predicción de mapas de contacto y los modelos de predicción de estructuras terciarias de extremo a extremo. Finalmente, se examina la posible dirección futura del aprendizaje profundo en el campo de la bioinformática estructural.

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

DESARROLLO

1. Arquitecturas de redes neuronales profundas

Las redes neuronales generalmente apilan múltiples módulos/capas. Cada uno de estos módulos podría modelarse como una función no lineal simple. En teoría, una red densa multicapa (es decir, MLP) puede aproximarse a cualquier función continua (Hornik,1989). Sin embargo, esto es difícil de lograr en aplicaciones del mundo real porque los datos son limitados y las funciones de destino son complicadas. Por lo tanto, varias arquitecturas de redes neuronales más efectivas con un sesgo inductivo más fuerte están diseñadas para procesar datos que tienen un formato específico. Para los datos de secuencias de proteínas, las redes neuronales convolucionales (CNN), las redes neuronales recurrentes (RNN) y las redes neuronales basadas en la atención se encuentran entre las arquitecturas más populares. Los presentamos brevemente en esta sección.

1.1. Redes neuronales convolucionales

La arquitectura de las CNN está relacionada con el Neocognitron y la red neuronal de retardo de tiempo (TDNN) (Fukushima, 1980). Las CNN modernas utilizan la retropropagación para la optimización de la red (LeCun, 1989). Las CNN están diseñadas para datos con estructura, como secuencias (datos 1D), imágenes (datos 2D) y nubes de puntos (datos 3D). Por ejemplo, cualquier píxel en una imagen está más correlacionado con sus píxeles vecinos espacialmente y los píxeles en la misma vecindad pueden formar un patrón local. Estos patrones no están relacionados con sus ubicaciones en la imagen y pueden formar patrones más grandes o de mayor nivel. Las capas convolucionales y las capas de agrupación en las CNN captaron patrones de bajo y alto nivel más fácilmente que las MLP. La implementación de la convolución y la agrupación es similar a una ventana deslizante o un filtro. La diferencia entre ellos es la función para procesar los datos en la ventana. Tomando el procesamiento de imágenes como ejemplo, cuando un lote de imágenes que se representan como un tensor $I \in R^{(N*W*H*D_{in})}$ son entrada en una capa convolucional, el tensor de salida $O \in R^{(N*W*H*D_{out})}$ puede ser derivado de la siguiente manera:

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

$$O_{i,j} = \sum_{d=1}^{D_{in}} W_d * I_{i,d} + b_j \qquad (1)$$

donde N denota el tamaño del lote, W y H denotan el tamaño de las imágenes, D_{in} y D_{out} denotan el número de canales de entrada y salida respectivamente, $i_{i,d}$denota el canal d^{th} del i^{th}imagen de entrada, y $O_{i,j}$ denota el j^{th}Canal de la i-ésima imagen de salida. $W_d$es la d-ésima matriz de peso aprendible, b_j es el j-ésimo sesgo aprendible y * Es el operador de correlación cruzada 2D válido. La operación de agrupación sustituye la función de suma ponderada con una función de máximo o promedio y, por lo general, tiene un paso mayor. Las CNN apilan múltiples capas convolucionales y capas de agrupación para aprender funciones. Cuando la red es más profunda, su capacidad crece, pero su rendimiento puede degradarse debido a problemas, como el sobreajuste y la desaparición de gradientes. Se propone una variedad de métodos para mitigar estos problemas, como abandono, normalización por lotes y mapa idéntico (Srivastava, 2014) (Ioffe, 2015). Estas técnicas también se utilizan en otros tipos de redes neuronales profundas.

1.2. Redes neuronales recurrentes

Los RNN son marcos efectivos para procesar datos de secuencia. En comparación con los MLP, los RNN tienen una celda de memoria para registrar el estado del paso del tiempo anterior. Un RNN simple se puede formular de la siguiente manera (Elman, 1990):

$$h_t = \sigma_h(W_h x_t + U h_{t-1} + b_h) \qquad (2)$$

$$y_t = \sigma_y(W_y h_t + b_y) \qquad (3)$$

Donde x_t, y_t y h_tEs la incrustación de entrada, la incrustación de salida, y vector de estado oculto en el paso de tiempo t, respectivamente. W, U y b son parámetros aprendibles y σ es la función de activación. Cuando las RNN se desarrollan a lo largo de secuencias largas, pueden considerarse redes neuronales profundas. Por lo tanto, las RNN también sufren la desaparición del gradiente. Para resolver este problema Para resolver el problema, se proponen las redes neuronales de memoria a corto plazo (LSTM), que añaden una célula de memoria e implementan el control de puerta (Hochreiter, 1997). Para cualquier vector de

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

entrada x_t en el paso t^{th},el vector de salida h_t (es decir, el vector de estado oculto) y la célula de memoria c_t se obtienen como sigue:

$$f_t = \sigma_g\left(W_f x_t + U_f h_{t-1} + b_f\right) \quad (4)$$

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \quad (5)$$

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \quad (6)$$

$$\tilde{c}_t = \sigma_g(W_c x_t + U_c h_{t-1} + b_c) \quad (7)$$

$$c_t = f_1 \,^\circ c_{t-1} + i_t{}^\circ \tilde{c}_t) \quad (8)$$

$$h_t = o_t \,^\circ \sigma_t(c_t) \quad (9)$$

donde f_t, i_t, o_t y $\tilde{c}_t$ son puerta de olvido, puerta de entrada, puerta de salida y entrada de célula respectivamente. W, U y b son parámetros aprendibles. σ_g, σ_c y σ_t son funciones sigmoideas, *tanh* y *tanh*, respectivamente. El operador ∘ es el producto elemento-sabio. LSTM aprende la dependencia a largo plazo porque las células de memoria pueden almacenar información de muchos pasos temporales anteriores. Ha habido varias aplicaciones exitosas de las LSTMs, como el reconocimiento del habla, la traducción automática, el control de robots y la predicción de la estructura y función secundaria de las proteínas (Berner, 2019).

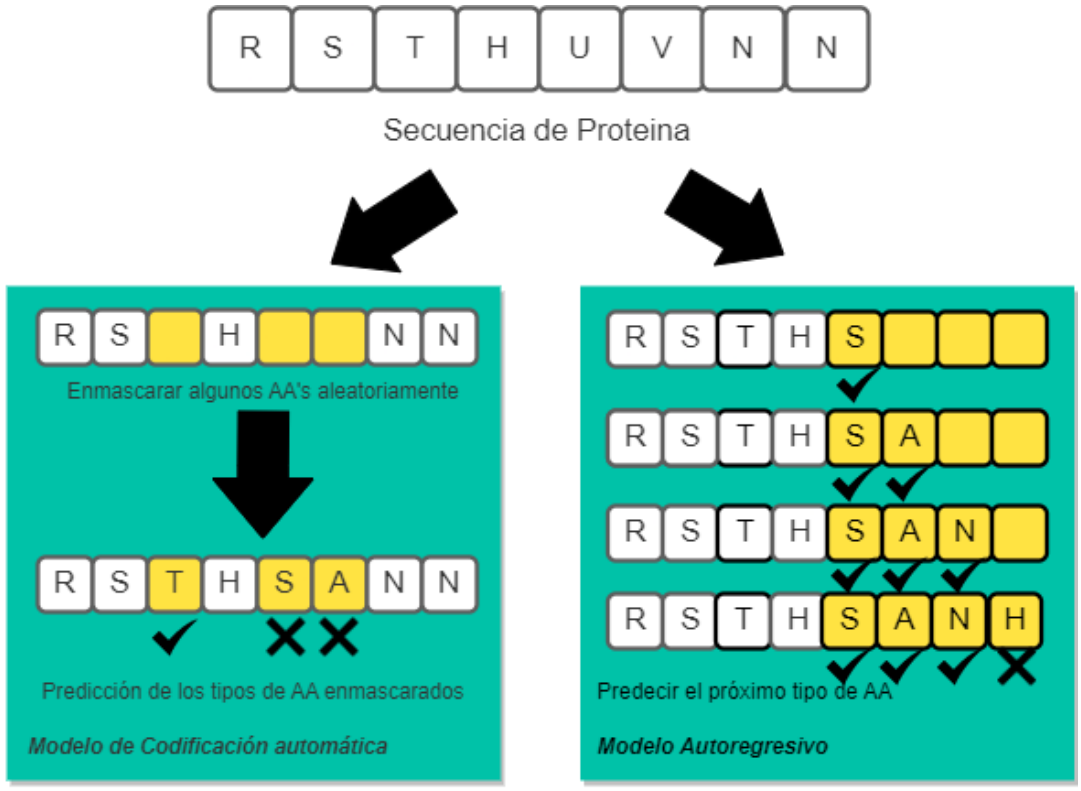
2. Representación de secuencias de proteínas basadas en secuencias únicas

Los modelos basados en una única secuencia aprenden la representación de proteínas a partir de una única codificación de la secuencia de proteínas. Los modelos lingüísticos que utilizan se clasifican en modelos autorregresivos (AR) y modelos de auto-codificación (AE). Como se muestra en la Figura 2, las tareas previas para los modelos de lenguaje de proteínas AR y AE son la predicción del siguiente AA y la predicción del AA enmascarado, respectivamente.

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

Figura 2

Tareas de pretexto para el modelado del lenguaje de las proteínas Auto Regresivas (AR) y Auto Codificación (AE)



Nota: elaboración propia.

Para una secuencia de proteína dada, los modelos AE enmascararán algunos aminoácidos (AAs) aleatoriamente y luego predecirá los tipos de AA enmascarados de acuerdo con los AAs conocidos restantes. Los modelos Auto Regresivos (AR) predecirán el siguiente tipo de AA según todos los AA anteriores de forma iterativa.

Tanto las RNN como las redes neuronales basadas en la atención pueden servir como codificador de los modelos AR, mientras que estas últimas pueden utilizarse además para los modelos AE debido a sus diferentes tareas de pretexto. Inicialmente, los modelos AR acceden a los contextos desde una sola dirección durante el entrenamiento. Posteriormente, se han propuesto modelos modificados, como XLNet, para permitir que los modelos AR aprendan de contextos bidireccionales en PLN, pero todavía hay mucho margen para seguir mejorando

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

el modelado de estos métodos de los datos de secuencias de proteínas (Yang, 2019). Por el contrario, los modelos AE aprenden de contextos bidireccionales pero pueden sufrir una discrepancia pre entrenamiento-ajuste debido a la introducción del token [MASK]. En la Tabla 1 resumimos parte de los modelos representativos de lenguaje proteico autosupervisado para la representación de secuencias de proteínas. El método TAPE introduce el Transformador en la representación de secuencias de proteínas.

Tabla 1.

Algunos modelos de lenguaje proteico autosupervisado para la representación de secuencias proteicas

Método	Input	Modelo de Lenguaje	Codificador	Datos Entrenados
TAPE	Secuencia Única	Codificación Automática	Transformer	Pfam
SeqVec	Secuencia Única	Auto Regresivo	ELMo	UniRef50
UniRep	Secuencia Única	Auto Regresivo	mLSTM	UniRef50
ESM-lb	Secuencia Única	Codificación Automática	Transformer	UniParc
ESM-MSA-lb	MSA	Codificación Automática	Transformer	UniRef50

Nota: elaboración propia.

3. Representación de secuencias de proteínas basadas en MSA

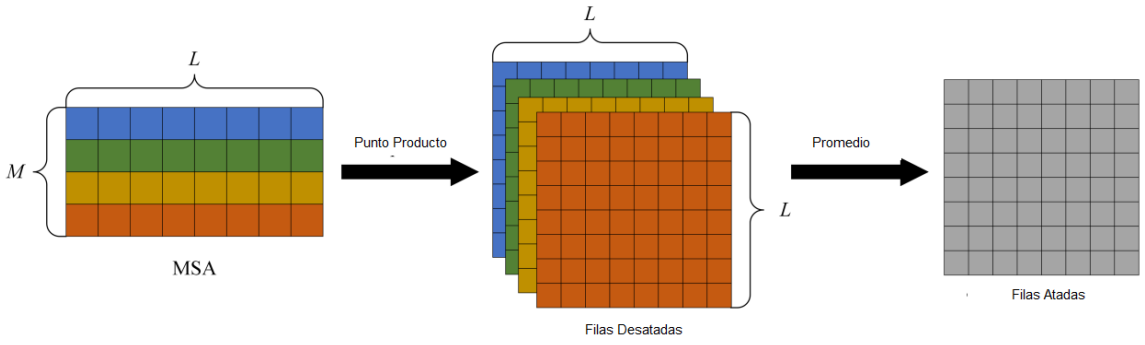
La entrada de un modelo basado en MSA suele ser el MSA de la secuencia en lugar de la propia secuencia. La tarea previa de un modelo basado en MSA es predecir AA enmascarados aleatoriamente en el MSA. Cabe señalar que los AA enmascarados pueden predecirse no sólo a partir de los residuos en el de la misma secuencia, sino también del residuo en la misma posición de otras secuencias en el MSA (Rao, 2019). Si simplemente concatenamos todas secuencias de un MSA, el tamaño de la matriz de atención será M2 ,L2 donde M es el número de secuencias en el MSA y L es la longitud de cada secuencia. La

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

complejidad espacial y temporal de la operación podría ser prohibitiva. En ESM-MSA-1b, se aplica una variante de la atención axial para reducir el coste computacional (Ho, 2019). Como se muestra en la Figura 3, la atención de fila desvinculada ayuda a reducir el coste de atención a O(ML2) y O(M2 L). Además, la atención hace que diferentes secuencias compartan la misma matriz de atención y reduce aún más el coste de atención a O(L2), con lo que se aprovecha al máximo la información evolutiva de la MSA.

Figura 3

Atención de fila atada y desatada utilizada en el ESM-MSA-1b



Nota: elaboración propia. M es el número de secuencias en el MSA y L es la longitud de cada secuencia

4. Predicción de estructura secundaria

La predicción de la estructura secundaria (SSP) es importante para la predicción de la estructura terciaria. Las estructuras secundarias se pueden utilizar para la búsqueda de plantillas en algoritmos de enhebrado (Gront, 2012) (Zheng, 2019). Los primeros métodos de SSP se basaban en y se desarrollaron en los años 70 (Chou, 1974). Estos métodos consideran principalmente los tipos de AA del residuo único objetivo o de un triplete centrado en él. Por lo tanto, la precisión Q3 (la fracción de residuos predichos correctamente para predicción de 3 clases) generalmente no es muy alta, es decir, ~60% (Torrisi, 2020). La estrategia de ventana deslizante La estrategia de ventana deslizante se utiliza entonces para extraer fragmentos continuos más largos como entrada, basándose en la suposición de que la estructura secundaria del residuo objetivo está muy relacionada con los residuos que lo

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

rodean. Además, a medida que se determinan más estructuras de proteínas, los potentes modelos de aprendizaje automático modelos de aprendizaje automático aplicados a SSP de la introducción de información evolutiva mejoran la precisión del Q3 hasta más del 70% (Torrisi, 2020). Las redes neuronales artificiales son actualmente, las redes neuronales artificiales son uno de los modelos de aprendizaje de aprendizaje automático para la SSP y mejoran aún más la precisión de la predicción.

4.1. Redes neuronales utilizadas para la predicción de estructuras locales

Los MLP (es decir, redes neuronales densas o redes neuronales totalmente conectadas) utilizados para SSP suelen contener una o dos capas ocultas. Los primeros enfoques basados en MLP aceptan como entrada una ventana deslizante de fragmentos de AA de codificación de un disparo y alcanzan una precisión de ~65% Q3 (Qian, 1988). Posteriormente, las características evolutivas fueron uno de los enfoques de codificación más populares. Por ejemplo, en PSIPRED, los PSSM normalizados se utilizaron como entrada del primer MLP y su precisión Q3 es del ~76% (Jones, 1999). Otros métodos similares también demostraron la eficacia de las características evolutivas (Rost, 1993; Cuff, 2000). Además de SSP, los MLP también se utilizan para otras propiedades locales de los esqueletos proteicos.

La versión inicial de SPIDER predice el ángulo θ (es decir, el ángulo formado por tres átomos Cα consecutivos) y el ángulo τ (es decir, el ángulo diedro formado por cuatro átomos Cα consecutivos). formado por cuatro átomos Cα consecutivos) basado en un MLP con tres capas ocultas (Lyons, 2014), cuya entrada consiste en las estructuras secundarias predichas y la superficie accesible al disolvente predicha a partir de SPINE-X (Faraggi, 2012). La versión SPIDER2 predice además estructuras secundarias, superficie accesible al disolvente y ángulos de torsión (Heffernan, 2015).

En general, los MLP utilizan la estrategia de ventana deslizante para derivar entradas de longitud fija. El tamaño de la ventana deslizante es un hiper parámetro importante y necesita un ajuste cuidadoso (Chen, 2006). Las ventanas deslizantes demasiado grandes pueden llevar a un sobreajuste o introducir ruido, mientras que las ventanas deslizantes demasiado pequeñas pueden perder información útil. Las RNN están diseñadas para datos secuenciales y pueden aceptar secuencias de longitud variable. Los modelos basados en RNN

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

suelen utilizar RNN bidireccionales para extraer el contexto de ambos lados del residuo objetivo. SSpro y su versión mejorada Porter, que alcanzan una precisión de ~79% Q3, se componen de un conjunto de RNN bidireccionales de dos etapas (BRNNs) (Baldi, 1999; Pollastri, 2005). SPIDER3 cambia su arquitectura a LSTMs bidireccionales (BiLSTMs) desde MLPs y alcanza una precisión de ~84% Q3 (Wang, 2016). Las CNN se suelen utilizar para datos 2D, como imágenes, pero también se han aplicado a SSP y su rendimiento es bueno. SSP y funcionan tan bien como otras arquitecturas de redes neuronales (DeepCNF sustituye los MLP de campos aleatorios condicionales (CRF) con CNNs que pueden capturar relaciones complejas entre características de entrada y las etiquetas de salida. Los experimentos muestran que la mejora de DeepCNF se debe principalmente a las CNN profundas, que también se han aplicado en otros modelos (Wang, 2016).

5. Los enfoques de SSP de última generación se benefician de datos más grandes, redes más profundas y mejores características evolutivas

Al investigar la diferencia entre algunos enfoques no consensuados de SSP y sus versiones actualizadas, los resultados resumidos en la Tabla 2 sugieren que la mejora de la SSP puede estar relacionada con varios factores. Teniendo en cuenta que esta tabla pretende comparar diferentes versiones del mismo método no métodos diferentes entre sí. Tras introducir el perfil de secuencia de PSI-BLAST, HHblits y MMseqs2 (Steinegger, 2017), la mayoría de los métodos de basados en el aprendizaje profundo pueden alcanzar más del 80% de precisión Q3 (Zhang, 2011). A continuación, estos enfoques utilizan más datos anotados para el entrenamiento y cambian.

Tabla 2.

Comparación entre PSI-PRED, Porter, SPIDER, NetSurfP y sus correspondientes versiones actualizadas.

Método	Número de Entrenamientos	Arquitectura	Codificador	Datos Entrenados
PSIPRED	~1100b	1-layer MLPs[c]	PSSM	76%
PSIPRED4	>1100b	2-layer MLPs	PSSM	84.2%
Porter	2171	BRNNs	PSSM	79%

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

Porter4.0	7522	BRNNs	PSSM	82.2%
SPIDER2	4590	3-layer MLPs	PSSM,PP[d]	82%
SPIDER3	4590	BiLSTMs	PSSM, PP, HMM[e]	84%
NetSurfP1.0	2085	1-layer MLPs	PSSM	81%
NetSurfP2.0	10337	HNNs	HMM, MMseqs2, onehot	85%

Nota: elaboración propia.

MLP poco profundas a redes neuronales más profundas, incluyendo MLP, RNN y redes neuronales híbridas (HNN) que se componen de diferentes tipos de capas. Por ejemplo, Netsurf2.0 implementa una HNN muy profunda que se compone de CNNs y RNNs y utiliza más de 10k secuencias de proteínas para el entrenamiento (Klausen, 2019). Consigue una mejora del 4% en la precisión Q3 en comparación con su versión MLP de una sola capa. Además, las mejores herramientas de MSA y la rápida acumulación de datos de secuencias de proteínas no anotadas también hacen que los perfiles de secuencias sean más precisos. más precisa. Al sustituir el perfil HMM utilizado en Netsurf2.0 por la representación del ESM-MSA-1b, la precisión Q8 (la fracción de residuos predichos correctamente para la predicción de 8 clases) aumenta en un 2% (Buchan, 2019). Los resultados indican que unas mejores características evolutivas de las redes neuronales auto supervisadas pueden ayudar a mejorar la SSP.

6. Predicción del mapa de contacto

Una de las aplicaciones más recientes y exitosas del aprendizaje profundo a la bioinformática estructural de proteínas es la predicción del mapa de contacto/distancia. Cuando los mapas de contacto/distancia se determinan con precisión, las estructuras de proteínas se pueden reconstruir directa o indirectamente con mejoras (por ejemplo, como un término de energía para algoritmos heurísticos) (Vassura, 2008; Nugent, 2012)

6.1. Redes neuronales utilizadas para la predicción de mapas de contacto

El principal objetivo de la predicción de mapas de contacto (CMP) es averiguar si cada par de residuos de una proteína está en contacto o no. La definición de contacto, que

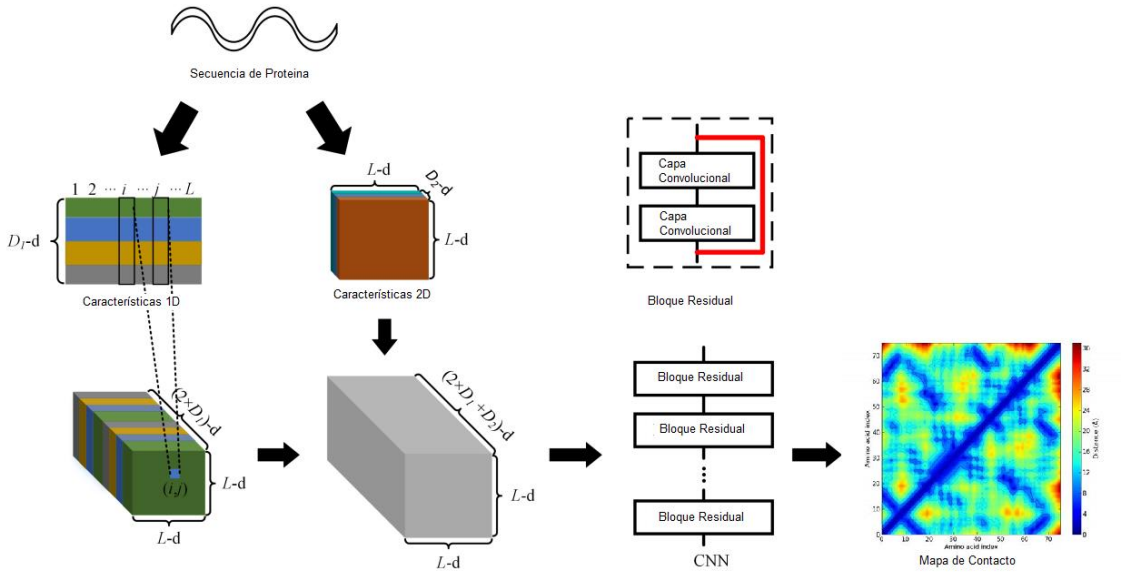
ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

también acepta el CASP, es que la distancia entre el Cβ de dos residuos AA (Cα para la Glicina) es inferior a 8Å. CMP pueden clasificarse generalmente en dos tipos: modelos estadísticos y modelos de aprendizaje automático. Los métodos basados en la covarianza fueron uno de los primeros modelos estadísticos de éxito. Se basan en la premisa biológica de que cuando un residuo AA muta, el residuo con el que interactúa simultáneamente para mantener la estructura y la función de la proteína correspondiente (Kortemme, 2004). Así, dos residuos que interactúan coevolucionan y sus AA están altamente correlacionados. La correlación entre un par de residuos puede derivarse aproximadamente del MSA calculando la información mutua (MI) y el coeficiente de correlación de Pearson (PCC) (Gobel, 1994; Gloor, 2005). DCA es otro tipo de método estadístico utilizado para evitar introducir el ruido transitivo implicado en los métodos de covarianza tradicionales (Rives, 2021), que puede considerarse como los métodos CMP no supervisados. Los métodos supervisados de aprendizaje automático suelen basarse en la salida de DCA y los perfiles de secuencia. La mayoría de los primeros métodos son modelos superficiales como SVM y bosques aleatorios (Cheng, 2007; Li, 2011). Aunque los MLP también se han aplicado a CMP, sus arquitecturas son poco adecuadas para procesar datos coevolutivos 2D y son relativamente superficiales. Por ejemplo, MetaPSICOV es un MLP de dos etapas con una capa oculta (Jones, 2015), que extrae el contexto de un par de residuos mediante una estrategia de ventana deslizante debido a la arquitectura de los MLP. En cambio, las CNN están diseñadas para datos 2D, como imágenes, y pueden ser transferidos para predecir CMP de proteínas tratando las características coevolutivas como imágenes (Yang, 2018). Los núcleos convolucionales son similares a las ventanas y pueden extraer características multiescala apilando múltiples capas convolucionales. En la Figura 4 se muestra el proceso básico de los métodos CMP basados en CNN. El uso de redes neuronales más profundas o el aprendizaje conjunto ayuda a aumentar la capacidad de los modelos. Por ejemplo, DNCON2 utilizó un conjunto de cinco CNN con diferentes umbrales de distancia y SPOT-Contact combina ResNet y BiLSTMs (Hanson, 2018; Adhikari, 2018).

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

Figura 4

Proceso básico de predicción de mapas de contacto de proteínas basado en CNN,



Nota: elaboración propia.

Las características 1D (por ejemplo, PSSM) y 2D (por ejemplo, características coevolutivas) se extraen de una secuencia de proteínas de longitud L. El número de tipos de características 1D y 1D es D1 y D2, respectivamente. de características 1D y características 1D es D1 y D2, respectivamente. A continuación, las características 1D se concatenan en 2D y se vuelven a concatenar con características 2D. Por último, se crea una red neuronal residual (es decir, una CNN con mapas idénticos) para predecir el mapa de contacto a partir de las características concatenadas de entrada. basándose en las características concatenadas de entrada.

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

DISCUSIÓN Y CONCLUSIONES

La predicción de la estructura de las proteínas ha experimentado un notable avance gracias a las técnicas de aprendizaje profundo. El éxito de estas técnicas está estrechamente relacionado con varios factores, como la rápida acumulación de datos de secuencias y arquitecturas de redes eficaces. En particular, se han desarrollado diversas arquitecturas de red para la predicción de la estructura de proteínas, como las redes neuronales convolucionales (CNN), las redes neuronales recurrentes (RNN) y las redes más recientes basadas en la atención, como BERT y Transformers.

Para entrenar estos grandes modelos, se han utilizado técnicas de aprendizaje y aumento de datos. Sin embargo, existen aún desafíos por resolver en este campo, como la presencia de choques estéricos en las estructuras predichas que requieren refinamiento, así como la falta de interpretabilidad en las principales redes neuronales profundas actuales, las cuales funcionan como cajas negras.

Es necesario realizar mayores esfuerzos para abordar estos problemas, ya que se espera que muchas tareas futuras se beneficien de estructuras de proteínas predichas con alta precisión. Por ejemplo, el acoplamiento de proteínas, la predicción del sitio de unión proteína-ligando y la predicción de la función de las proteínas son solo algunas de las diversas aplicaciones que pueden mejorar nuestra comprensión general de las estructuras y sus complejas funciones. En consecuencia, se requiere una investigación continua en este campo para mejorar la calidad y precisión de las predicciones de estructura de proteínas y, por ende, avanzar en nuestra comprensión del mundo molecular.

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

REFERENCIAS BIBLIOGRÁFICAS

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, *181*(4096), 223-230.

Porter, L. L. y Looger, L. L. (2018). Extant fold-switching proteins are widespread. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(23), 5968-5973.

Pauwels, K., Van Molle, I., Tommasen, J. y Van Gelder, P. (2007). Chaperoning Anfinsen: The steric foldases. *Molecular Microbiology*, *64*(4), 917-922.

Oldfield, C. J., Uversky, V. N., Dunker, A. K. y Kurgan, L. (2019). Introduction to intrinsically disordered proteins and regions. N. Salvi (ed.). *Intrinsically Disordered Proteins*. Academic Press, (pp. 1-34).

Lieutaud, P., Ferron, F., Uversky, A. V., Kurgan, L., Uversky, V. N. y Longhi, S. (2016). How disordered is my protein and what is its disorder for? A guide through the "dark side" of the protein universe. *Intrinsically Disordered Proteins*, *4*(1), e1259708.

Zhang Y. (2008). Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, *18*(3), 342-348.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J. y Zhang, Y. (2015). The I-TASSER suite: Protein structure and function prediction. *Nature Methods*, *12*(1), 7-8.

Pearce, R. y Zhang, Y. (2021). Toward the solution of the protein structure prediction problem. *Journal of Biological Chemistry*, *297*(1), 100870.

Bengio, Y., Courville, A. y Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828.

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Van Esesn, B. C., Awwal, A. A. S. y Asari, V. K. (2018). *The history began from*

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

AlexNet: A comprehensive survey on deep learning approaches. arXiv. https://doi.org/10.48550/arXiv.1803.01164

Krizhevsky, A., Sutskever, I. y Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *International Conference on Neural Information Processing Systems*, Lake Tahoe, Nevada. Curran Associates Inc., 1097-1105.

Amodei D, et al. (2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. *Proceedings of The 33rd International Conference on Machine Learning*, M. F. Balcan y K. Q. Weinberger (eds.), PMLR, 173-182.

Alipanahi, B., Delong, A., Weirauch, M. T. y Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNAbinding proteins by deep learning. *Nature Biotechnology*, *33*(8), 831-838.

AlQuraishi M. (2019). End-to-end differentiable learning of protein structure. *Cell Systems*, *8*(4), 292-301.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J. y Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(15), 1-12.

Weigt, M., White, R. A. Szurmant, H., Hoch, J. A. y Hwa, T. (2009). Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(1), 67-72.

Jumper J, et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *96*(7873), 583-589.

Hornik, K., Stinchcombe, M. y White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359-366.

Fukushima K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*, 193-202.

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

LeCun Y, Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. y Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*, 541-551.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. y Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929-1958.

Ioffe, S. y Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *23 International Conference on Machine Learning*, B. Francis y B. David (eds.), PMLR, (pp. 448-456).

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 213-252.

Hochreiter, S. y Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780.

Berner, C., et al. (2019). *Dota 2 with large scale deep reinforcement learning*. arXiv.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R. y Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. *Proceedings of the 33rd International Conference on Neural Information Processing Systems, NeurIPS*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox y R. Garnett (eds.). Curran Associates. (pp. 5754-5764).

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P. y Song, Y. (2019). Evaluating protein transfer learning with TAPE. *Proceedings of the 33rd International Conference on Neural Information Processing Systems, NeurIPS*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox y R. Garnett (eds.). Curran Associates. (pp. 9686-9698).

Gront, D., Blaszcyk, M., Wojciechowski, P. y Kolinski, A. (2012). BioShell threader: Protein homology detection based on sequence profiles and secondary structure profiles. *Nucleic Acids Research*, *40*, 257-262.

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

Zheng, W., Zhang, C., Wuyung, Q., Pearce, R., Li, Y. y Zhang, Y. (2019). LOMETS2: Improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins. *Nucleic Acids Research*, *47*(1), 429-436.

Chou, P. Y. y Fasman, G. D. (1974). Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry*, *13*(2), 211-222.

Torrisi, M., Pollastri, G. y Le, Q. (2020). Deep learning methods in protein structure prediction. *Computational and Structural Biotechnology Journal*, *18*, 1301-1310.

Qian, N. y Sejnowski, T. (1988). Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology*, *202*(4), 865-884.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, *292*(2), 195-202.

Rost, B. y Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology*, *232*(2), 584-599.

Cuff, J. A. y Barton, G. J. (2000). Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, *40*(3), 502-511.

Lyons, J., Dehzangi, A., Heffernan, R., Sharma, A., Paliwal, K., Sattar, A., Zhou, Y. y Yang, Y. (2014). Predicting backbone Calpha angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. *Journal of Computational Chemistry*, *35*(28), 2040-2046.

Faraggi, E., Zhang, T., Yang, Y., Kurgan, L. y Zhou, Y. (2012). SPINE X: Improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *Journal of Computational Chemistry*, *33*(3), 259-267.

Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., Yang, Y. y Zhou, Y. (2015). Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Scientific Reports*, *5*, 11476.

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

Chen, K., Kurgan, L. y Ruan, J. (2006). Optimization of the sliding window size for protein structure prediction. *Proceedings of the 2006 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, Toronto, IEEE.

Baldi, P., Brunak, S., Frasconi, P., Soda, G. y Pollastri, G. (1999). Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics, 15*(11), 937-946.

Pollastri, G. y McLysaght, A. (2005). Porter: A new, accurate server for protein secondary structure prediction. *Bioinformatics, 21*(8), 1719-1720.

Wang, S., Peng, J., Ma, J. y Xu, J. (2016). Protein secondary structure prediction using deep convolutional neural fields. *Scientific Reports, 6*, 18962.

Wang, S., Li, W., Liu, S. y Xu, J. (2016). RaptorX-Property: A web server for protein structure property prediction. *Nucleic Acids Research*, *44*(1), 430-435.

Steinegger, M. y Soding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, *35*(11), 1026-1028.

Zhang, H., Zhang, T., Chen, K., Kedarisetti, K. D. Mizianty, M. J., Bao, Q., Stach, W. y Kurgan, L. (2011). Critical assessment of high-throughput standalone methods for secondary structure prediction. *Briefings in Bioinformatics*, *12*(6), 672-688.

Klausen, M. S., et al. (2019). NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins-Structure Function and Bioinformatics, 87*(6), 520-527.

Buchan, D. W. A. y Jones, D. T. (2019). The PSIPRED protein analysis workbench: 20 years on. *Nucleic Acids Research, 47*(1), 402-407.

Vassura, M., Margara, L., Di Lena, P., Medri, F., Fariselli, P. y Casadio, R. (2008). Reconstruction of 3D structures from protein contact maps. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(3), 357-367.

Nugent, T. y Jones, D. T. (2012). Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s

analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(24), 1540-1547.

Kortemme, T., Joachimiak, L. A., Bullock, A. N., Schuler, A. D., Stoddard, B. L. y Baker, D. (2004). Computational redesign of protein-protein interaction specificity. *Nature Structural & Molecular Biology*, *11*(4), 371-379.

Gobel, U., Sander, C., Schneider, R. y Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins*, *18*(4), 309-317.

Gloor, G. B., Martin, L. C., Wahl, L. M. y Dunn, S. D. (2005). Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry*, *44*(19), 7156-7165.

Cheng, J. L. y Baldi, P. (2007). Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, *8*, 113.

Li, Y. Q., Fang, Y. y Fang, J. (2011). Predicting residue-residue contacts using random forest models. *Bioinformatics*, 27(24), 3379-3384.

Jones, D. T., Singh, T., Kosciolek, T. y Tetcher, S. (2015). MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long-range hydrogen bonding in proteins. *Bioinformatics*, *31*(7), 999-1006.

Yang, J. y Shen, H. B. (2018). MemBrain-contact 2.0: A new two-stage machine learning model for the prediction enhancement of transmembrane protein residue contacts in the full chain. *Bioinformatics*, *34*(2), 230-238.

Hanson, J., Paliwal, K., Litfin, T., Yang, Y. y Zhou, Y. (2018). Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics*, *34*(23), 4039-4045.

Adhikari, B., Hou, J. y Cheng, J. (2018). DNCON2: Improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics*, *34*(9),1466-1472.

ISBN: 978-628-95884-1-5 - DOI: https://doi.org/10.34893/e1150-3660-8721-s