13. PREDICCIÓN DE INTERACCIONES DE PROTEÍNAS CON MACHINE LEARNING: UNA HERRAMIENTA AVANZADA PARA LA INVESTIGACIÓN BIOMOLECULAR

Predicting Protein Interactions with Machine Learning: An Advanced Tool for Biomolecular Research

Jordan Piero Borda Colque[77]

Bernabé Canqui Flores [78]

Alfredo Tumi Figueroa [79]

Fred Torres-Cruz [80]

Juan Kenyhy Hancco Quispe[81]

Pares evaluadores: Red de Investigación en Educación, Empresa y Sociedad – REDIEES.[82]

[77] Universidad Nacional del Altiplano de Puno, P.O. Box 291, Puno-Peru, https://orcid.org/0000-0001-8488-1658 , jbordac@est.unap.edu.pe

[78] Universidad Nacional del Altiplano de Puno, P.O. Box 291, Puno-Peru, https://orcid.org/0000-0003-2204-0620, bcanqui@unap.edu.pe

[79] Facultad de Medicina Humana, Universidad Nacional del Altiplano de Puno, P.O. Box 291, Puno-Perú, https://orcid.org/0000-0003-2970-061X, alfredo2891@yahoo.es

[80] Universidad Nacional del Altiplano de Puno, P.O. Box 291, Puno-Peru, https://orcid.org/0000-0003-0834-6834 , ftorres@unap.edu.pe

[81] Universidad Nacional del Altiplano de Puno, P.O. Box 291, Puno-Peru, https://orcid.org/0000-0002-2125-0530, jhanccoq@est.unap.edu.pe

[82] 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

PREDICCIÓN DE INTERACCIONES DE PROTEÍNAS CON APRENDIZAJE AUTOMÁTICO: UNA HERRAMIENTA AVANZADA PARA LA INVESTIGACIÓN BIOMOLECULAR

Jordan Piero Borda Colque, Bernabé Canqui Flores, Alfredo Tumi Figueroa, Fred Torres-Cruz, Juan Kenyhy Hancco Quispe

RESUMEN

Las proteínas, como biomoléculas esenciales para la vida, participan en una variedad de procesos celulares. Son moléculas de gran tamaño que constan de uno o más polipéptidos, los cuales se pliegan en una estructura tridimensional única que determina su función biológica. La determinación de la estructura tridimensional de las proteínas es esencial para comprender sus funciones. En el campo de la bioinformática, el desarrollo de enfoques computacionales que permitan predecir la estructura de las proteínas a partir de su secuencia de aminoácidos es una de las áreas más activas de investigación. En este contexto, la predicción de aspectos topológicos clave de la estructura, como los contactos y distancias entre residuos de proteínas, se ha convertido en una herramienta importante para la predicción de la estructura tridimensional de proteínas basada en secuencia. En este capítulo, presentamos una revisión sistemática de los métodos más representativos para la predicción de la interacción geométrica entre residuos de proteínas. Entre ellos, se incluyen los métodos basados en correlación, los métodos de análisis de acoplamiento directo y sus estrategias de posprocesamiento, los métodos clásicos de aprendizaje automático y los métodos avanzados de aprendizaje profundo. Además, se discuten las aplicaciones de estas interacciones en la predicción de la estructura tridimensional de las proteínas, que es fundamental para comprender la función biológica de estas moléculas.

Palabras Clave: biomoléculas; polipéptidos; bioinformática; aprendizaje automático; aprendizaje profundo.

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

ABSTRACT

Proteins, as essential biomolecules for life, participate in a variety of cellular processes. They are large molecules that consist of one or more polypeptides, which fold into a unique three-dimensional structure that determines their biological function. The determination of the three-dimensional structure of proteins is essential to understand their functions. In the field of bioinformatics, the development of computational approaches that make it possible to predict the structure of proteins from their amino acid sequence is one of the most active areas of research. In this context, the prediction of key topological aspects of structure, such as contacts and distances between protein residues, has become an important tool for sequence-based three-dimensional protein structure prediction. In this chapter, we present a systematic review of the most representative methods for the prediction of the geometric interaction between protein residues. These include correlation-based methods, tightly coupled analysis methods and their post-processing strategies, classical machine learning methods, and advanced deep learning methods. In addition, the applications of these interactions in the prediction of the three-dimensional structure of proteins are discussed, which is essential to understand the biological function of these molecules.

Keywords: biomolecules; polypeptides; bioinformatics; machine learning; deep learning.

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

INTRODUCCIÓN

Las proteínas son componentes esenciales de las células vivas y representan uno de los elementos básicos más importantes de la vida. Estas biomoléculas se forman a través de la deshidratación y condensación de una secuencia de aminoácidos. Aunque la secuenciación de proteínas puede ser relativamente fácil de obtener a través de técnicas de alto rendimiento, determinar su conformación tridimensional (3D) es una tarea compleja. Los métodos experimentales para la determinación de la estructura incluyen el análisis de rayos X (Kendrew,1958), la resonancia magnética nuclear (RMN) (Wüthrich, 2001), y la técnica emergente de microscopía crioelectrónica (Taylor, 1974).

En la era posgenómica, la determinación de la estructura 3D de las proteínas a través de experimentos está lejos de alcanzar la acumulación de una gran cantidad de secuencias de proteínas. Por lo tanto, el mejoramiento de la precisión en la predicción de la estructura 3D de las proteínas directamente a partir de sus secuencias de aminoácidos se ha convertido en un tema clave en el campo de la bioinformática estructural (Baker, 2001). Debido a las limitaciones de las técnicas experimentales, la predicción directa de la estructura de las proteínas a partir de su secuencia está ganando gradualmente atención.

Las investigaciones realizadas por Anfinsen (1973) demuestran que la secuencia de aminoácidos de las proteínas determina su estructura 3D, lo que sugiere que la información estructural de las proteínas puede integrarse completamente en sus secuencias. Esta hipótesis ha impulsado el desarrollo de la predicción de la estructura de proteínas basada en secuencias. Los enfoques convencionales para la predicción de la estructura de proteínas, por ejemplo, el modelado de homología (Webb, 2016; Schwede, 2003) y el enhebrado (Bowie, 1991), requieren una plantilla de estructura razonablemente similar para estar disponible en una base de datos de proteínas existente. Los métodos de predicción ab initio se están volviendo cada vez más importantes ya que las plantillas son innecesarias para estos métodos. Por lo general, los métodos de predicción de estructuras ab initio se aplican tanto a la física como a los potenciales basados en el conocimiento para guiar la simulación del plegamiento.

La predicción de las coordenadas de los átomos que componen la estructura de la proteína directamente a partir de la secuencia correspondiente mediante el aprendizaje

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

automático puede parecer natural para los desarrolladores en el campo. Sin embargo, las coordenadas son difíciles de predecir ya que cambiarán con las transformaciones de rotación y traslación, mientras que la estructura misma permanecerá sin cambios. Alternativamente, algunas representaciones invariables de la estructura de la proteína, por ejemplo, contactos entre residuos, se pueden predecir utilizando métodos de aprendizaje automático. Los patrones de los mapas de contacto de proteínas, donde cada entrada indica si el par de residuos correspondiente está en contacto, reflejan algunas de las características clave de las estructuras.

Más importante aún, las representaciones invariantes predichas se pueden usar como potenciales para construir un campo de fuerza más confiable para la predicción de la estructura de la proteína. Se han hecho progresos considerables para mejorar el rendimiento de las predicciones del mapa de contactos de proteínas en los últimos años. Los enfoques iniciales se centran en analizar las correlaciones por pares para abordar los contactos físicos basados en modelos de aprendizaje automático no supervisados (Cocco, 2018). Los modelos de aprendizaje automático supervisados se aplicaron aún más para aprender información a través de alineaciones de secuencias múltiples (MSA) de datos de entrenamiento (Cheng, 2007). Además de la evolución de los modelos de aprendizaje automático, los términos de geometría invariable también están evolucionando de contactos a distancia y orientaciones (Senior, 2020). Este capítulo presentará una revisión sistemática de enfoques representativos en esas categorías.

MÉTODOS COMPUTACIONALES PARA LA PREDICCIÓN DE INTERACCIONES ENTRE RESIDUOS DE PROTEÍNAS

Definición de términos geométricos para la interacción entre residuos

La representación fundamental para modelar interacciones entre residuos de proteínas son los contactos entre sus átomos Cβ (Cα en el caso de glicina). Dos residuos se consideran en contacto si la distancia euclidiana entre sus átomos de Cβ es menor que un umbral predefinido, que suele ser de 8 Å, como se define en los experimentos de Evaluación crítica

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

de la predicción de la estructura de proteínas (CASP) (Li, 2021; Moult, 2005). Los contactos entre residuos se clasifican en tres categorías: corto, mediano y largo alcance, según la separación secuencial entre los residuos en contacto, que abarca desde 6 a 11, 12 a 23 y más de 24 posiciones, respectivamente. Los contactos de largo alcance tienen un mayor impacto en la estabilización de la topología de la proteína, por lo que se les da mayor importancia en la evaluación de la predicción de estructuras proteicas.

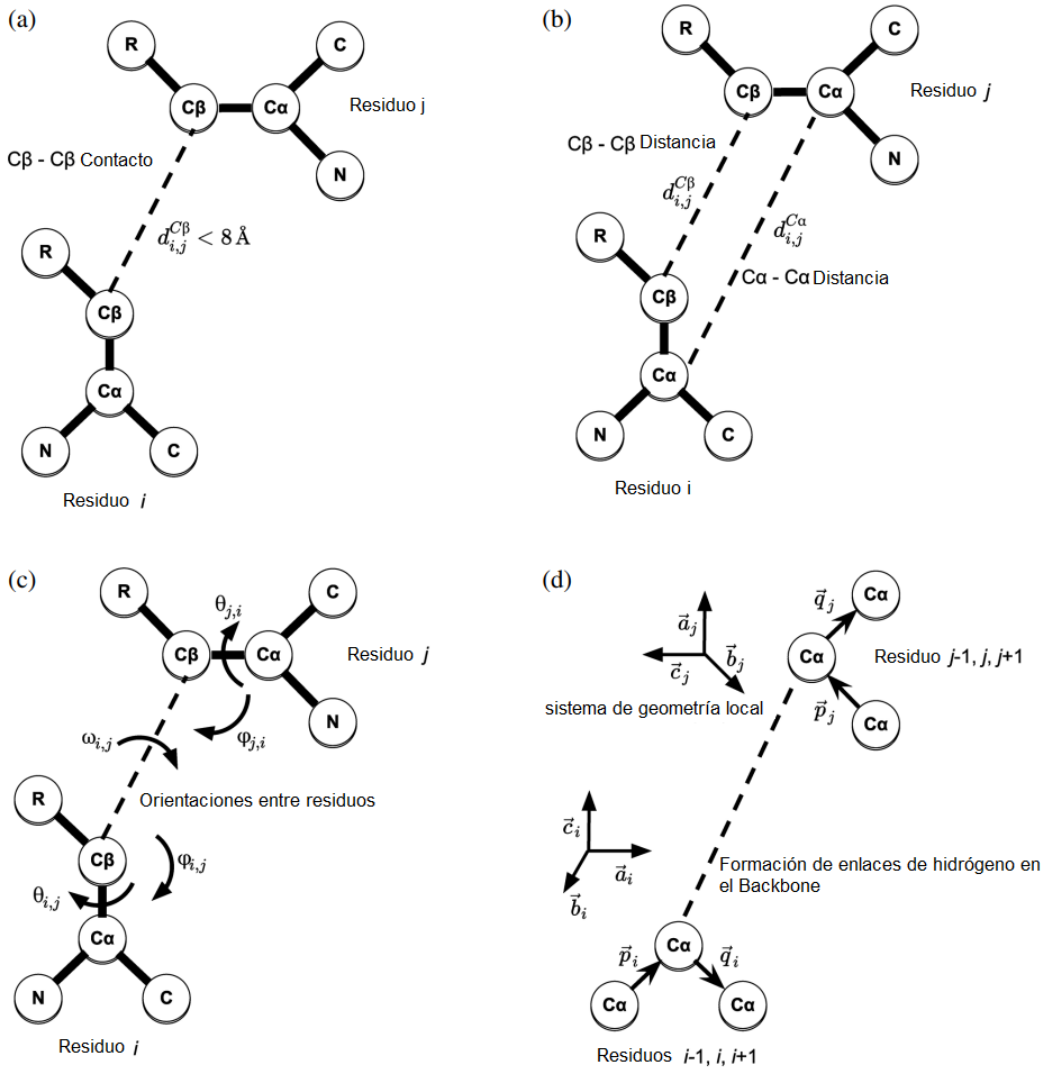
Además de los contactos binarios, se han desarrollado métodos para predecir la distancia discretizada entre dos átomos, como se implementa en AlphaFold (Senior AW,2020) en CASP13, lo que proporciona más información para el plegamiento de proteínas. Sin embargo, la información de distancia sola puede no ser suficiente para representar adecuadamente la estructura de proteínas, especialmente su quiralidad. Por lo tanto, se han propuesto métodos adicionales, como trRosetta (Yang, 2020), que también considera un conjunto de términos de orientación, incluidos los ángulos diedro ω y θ y el ángulo ϕ.

Recientemente, DeepPotential (Li, 2021) ha incorporado un conjunto adicional de términos geométricos entre residuos, que incluyen tres términos relacionados con enlaces de hidrógeno (enlace H) que involucran seis átomos de Cα. Dados dos residuos i y j, se construyen dos marcos locales a partir de sus átomos de Cα adyacentes, respectivamente, y se definen los términos del enlace H como los ángulos entre los ejes x, y y z correspondientes de los dos marcos construidos. La figura 1 muestra la definición de estos términos geométricos.

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

Figura 1

Definición de términos de geometría entre residuos: (a) contactos, (b) distancia, (c) orientaciones y (d) términos de enlaces H.



Nota: elaboración propia.

Métodos no supervisados para la predicción de mapas de contacto

Los métodos no supervisados para la predicción de contactos entre residuos obtienen información de la coevolución de pares de residuos mediante el análisis de alineaciones múltiples de secuencias (MSA, por sus siglas en inglés). El MSA representa una colección de secuencias homólogas relacionadas genéticamente, que proporcionan información

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

genética relevante para la secuencia de proteína de interés. En el análisis de un sitio específico en el MSA, se puede determinar el grado de conservación evolutiva de ese sitio. Los aminoácidos en diferentes posiciones no mutan de forma independiente y la mayoría de las mutaciones de un solo punto son perjudiciales y pueden alterar la compatibilidad física en el entorno del sitio de mutación. Las mutaciones compensatorias en sitios vecinos a menudo pueden reparar el daño causado por mutaciones de un solo punto, lo que se conoce como coevolución.

Existen dos enfoques principales para analizar los patrones coevolutivos: los métodos basados en la correlación local y los métodos basados en el Análisis de Acoplamiento Directo (DCA, por sus siglas en inglés). La Tabla 1 resume algunos métodos representativos no supervisados para inferir los contactos entre residuos. En la literatura, un MSA dado se denota como A, y se utilizan M y L para indicar el número de secuencias en el MSA y la longitud de la secuencia de consulta, respectivamente.

Tabla 1

Colección de métodos de análisis de coevolución representativos para inferir los contactos entre residuos de proteínas.

Tipo	Métodos	Año	Aproximación	Disponibilidad
Local	Gobel et al	1994	NA	NA
	Yanofsky et al	1964	NA	NA
	Korber et al.	1993	NA	NA
	Martin et al.	2005	NA	NA
	OMES	2002	NA	NA
	Lapedes et al.	1999	Monte Carlo	NA
	mpDCA	2009	Message passing	NA
	mfDCA	2011	Mean field	NA
Global	PSICOV	2011	Graphical lasso	https://github.com/psipred/psicov
	Ricmap	2019	Graphical ridge	https://zhanggroup.org/ResPRE/
	plmDCA	2013	Pseudolikelihood	https://github.com/magnusekeberg/plmDCA
	GREMLIN	2013	Pseudolikelihood	https://gremlin.bakerlab.org/
	CCMpred	2014	Pseudolikelihood	https://github.com/soedinglab/CCMpred

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

clmDCA	2018	Composite Likelihood	http://protein.ict.ac.cn/clmDCA-

Nota: elaboración propia.

En el análisis de la MSA, es común identificar el tipo de aminoácido presente en la posición l de la secuencia alineada m-ésima. Cada entrada en la MSA puede tener 21 estados, que incluyen 20 tipos de aminoácidos naturales y un estado adicional que representa la brecha. Para puntuar la coevolución entre un par de residuos, se utilizan diversos coeficientes de correlación locales. Por ejemplo, Gobel et al. (Göbel, 1994) emplearon el coeficiente de correlación de Pearson (PCC) para predecir los efectos de correlación de residuos. También se ha utilizado la información mutua (Korber, 1993) para medir la relación no lineal entre residuos en MSA, y el coeficiente Observed Minus Expected Squared (OMES) (Kass, 2002), que cuantifica la diferencia entre las frecuencias observadas y esperadas de coocurrencia de residuos en dos posiciones cualesquiera.

Sin embargo, estos métodos pueden no producir resultados satisfactorios debido a su incapacidad para modelar acoplamientos. Por ejemplo, si las posiciones i y j, y las posiciones j y k están correlacionadas, los coeficientes de correlación locales pueden mostrar una alta correlación entre la posición i y k, aunque es posible que no estén físicamente en contacto. Para abordar esta limitación, se han propuesto enfoques basados en el Análisis de Acoplamiento Directo (DCA) que utilizan el modelo de Potts (Wu, 1982) para modelar los datos de MSA y considerar los acoplamientos entre posiciones. Estos métodos, que se conocen como modelos globales, suelen utilizar el modelo de Potts generalizado.

$$H(a_1 \ldots . a_L) = \left(\sum_{i=1}^{L} h_i(a_i) + \sum_{i \neq j}^{L} J_{ij}(a_i, a_j) \right) \quad (1)$$

donde $H(a_1, \ldots, a_L)$ es el Hamiltoniano del sistema y cada posición puede considerarse como una partícula; cada secuencia alineada es una observación. $h_i(a_i)$ y $J_{ij}(a_i, a_j)$ son los parámetros de campo locales en la posición i con el tipo de residuo a_i y los parámetros de acoplamiento en las posiciones i y j con residuos a_i y a_j, respectivamente. En

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

consecuencia, se puede construir un modelo estadístico global $P(a_1, \ldots, a_L)$ sobre todo el MSA, en el que la probabilidad de una secuencia en el MSA se puede definir como:

$$P(a_1 \ldots . a_L) = \frac{\exp\{-H(a_1 \ldots . a_L)\}}{Z} = \frac{1}{z}\left(\sum_{i=1}^{L} h_i(a_i) + \sum_{i \neq j}^{L} J_{ij}(a_i, a_j)\right) \quad (2)$$

donde Z es la constante de normalización, es decir, la función de partición, que garantiza $\sum aP(a) = 1$. El modelo debe ser coherente con las estadísticas empíricas, es decir, $P(a_i) = f_i(a_i)\ y\ P(a_i, a_j) = f_{ij}(a_i, a_j)$. Aquí $f_i(a_i)\ y\ f_{ij}(a_i, a_j)$ son las frecuencias estadísticas de posiciones individuales y por parejas en el MSA, respectivamente. Además de las restricciones anteriores, el análisis de acoplamiento directo requiere que se obtengan los parámetros J y h minimizando la función de verosimilitud logarítmica negativa, como se muestra en la ecuación (3):

$$S = -\sum_{m=1}^{M} logP(a_1^m \ldots . a_L^m) \quad (3)$$

Dado que se requiere la suma de los términos 21^L en la función de semejanza para calcular la constante de normalización Z, que es computacionalmente intratable, se han desarrollado muchas aproximaciones y estrategias alternativas.

Lapedes et al. (Lapedes, 1999) introdujo por primera vez el modelo de Potts para estimar los parámetros de acoplamiento entre posiciones aproximando la función de partición con monte carlo (MC). Sin embargo, para el modelo de Potts de 21 estados, la convergencia solo se puede garantizar después de tiempos exponenciales. Diez años más tarde, Martin Weight introdujo formalmente el concepto de guerra DCA en 2009 (Weigt, 2009), y se utilizó un método basado en el paso de mensajes como herramienta aproximada, lo que proporciona una solución única.

Morcos et al. (Morcos, 2011) propuso usar campos medios para aproximar los parámetros del modelo de Potts, y la aproximación de campo medio estándar se puede usar para estimar de manera autoconsistente los valores de borde de un solo punto. El parámetro de acoplamiento J puede obtenerse invirtiendo la matriz de convarianza. PSICOV para predecir contactos de residuos de proteínas. PSICOV primero procesó el MSA transformado

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

por $C_{ij}(a,b) = f_{ij}(a,b) - f_i(a)f_j(b)$. De esta forma, el modelo de Potts original se puede aproximar como un modelo gaussiano de Markov amutivariable. La matriz inversa de la matriz de covarianza se puede obtener minimizando la función objetivo (4) siguiente:

$$tr(C\Theta) - indet(\Theta) + \lambda \sum |\Theta_y| \tag{4}$$

Donde C es la matriz de covarianza y Θ es la matriz simétrica definida positiva a obtener, generalmente llamada matriz de precisión. tr(X) es la traza de la matriz X. Se puede estimar el parámetro de acoplamiento J = –Θ. Los dos primeros términos se pueden interpretar como la verosimilitud logarítmica negativa de los datos distribuidos gaussianos multivariantes, y el tercer término es el término de regularización L1 de Θ. La aproximación de campo medio y la aproximación gaussiana son los esquemas computacionalmente más efectivos para resolver el problema inverso de Potts. Los resultados experimentales en 150 secuencias PFAM muestran que PSICOV supera el enfoque inicial de campo medio. Li et al. (Li, 2019) propuso Ricmap, que estima la matriz de precisión mediante la regularización de crestas. La función de pérdida para la estimación matricial de precisión se convierte en:

$$tr(C\Theta) - indet(\Theta) + \lambda \sum \Theta_y^2 \tag{5}$$

En tal caso, la solución de forma cerrada se puede obtener:

$$\Theta = QKQ^T \tag{6}$$

donde Q son los vectores propios de C y K es una matriz diagonal cuyos elementos diagonales son

$$K_{i,i} = \frac{-\Lambda_{i,i} + \sqrt{\Lambda_{i,i}^2 + 8\lambda}}{4\lambda} \tag{7}$$

Y donde $\Lambda_{i,i}$son los valores propios de C. Ricmap se comparó en un conjunto de datos compuesto por objetivos CASP11 y CASP12 y se informó que superó marginalmente a PSICOV (Li, 2019). Sin embargo, la aproximación gaussiana ignora la propiedad de los datos etiquetados en cada ubicación en el MSA y trata las variables etiquetadas como variables continúas utilizando el enfoque de codificación one-hot. La aproximación gaussiana y de campo medio no puede converger a la solución óptima incluso con el número infinito de

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

muestras. Magnus Ekeberg et al. (Ekeberg, 2013) propusieron una aproximación de pseudoverosimilitud para inferir el modelo de Potts. La pseudoverosimilitud aproxima la probabilidad de una sola secuencia como el producto de las probabilidades de las posiciones en una sola secuencia:

$$P(a_i \dots a_L) = \prod_{i=1}^{L} P(a_i) \tag{8}$$

donde la probabilidad de un solo sitio se puede aproximar como:

$$p(a_l) = \frac{\exp(h_l(a_l) + \sum_{k=1,h\neq l}^{L} J_{lk}(a_l, a_k))}{\sum_{q=1}^{21} \exp(h_l(q) + \sum_{k=1,h\neq l}^{L} J_{lk}(q, a_k))} \tag{9}$$

Para un MSA con secuencias M, la función de pérdida de entropía cruzada de la aproximación de pseudoverosimilitud es:

$$S_{pseudo} = -\sum_{m=1}^{M}\sum_{l=1}^{L} \log\left(\frac{\exp(h_l(a_l) + \sum_{k=1,h\neq l}^{L} J_{lk}(a_l, a_k))}{\sum_{q=1}^{21} \exp(h_l(q) + \sum_{k=1,h\neq l}^{L} J_{lk}(q, a_k))}\right) \tag{10}$$

Esta función de pérdida es convexa; por lo tanto, se puede optimizar utilizando L-BFGS (Liu DC,1989) o el método de gradiente conjugado (Hestenes, 1952). Hay varias implementaciones de métodos DCA basados en la aproximación de pseudoverosimilitud, como GREMLIN (Kamisetty, 2013), plmDCA (Ekeberg, 2013) y CCMpred (Seemayer, 2014). Las implementaciones anteriores utilizan términos de regularización L2 para evitar el sobreajuste. GREMLIN intenta introducir información previa en el DCA. Por lo tanto, su rendimiento es ligeramente mejor que el de las otras dos implementaciones, pero los tres métodos superan a mfDCA para la aproximación de campo medio y a PSICOV para la aproximación gaussiana (Kamisetty, 2013). Zhang et al. (Zhang, 2018) propuso el método clmDCA, que utilizó la maximización de probabilidad compuesta (CLM), para aproximarse mejor a la función de probabilidad original. A diferencia de la pseudoverosimilitud, la función de verosimilitud compuesta se define como:

$$\zeta_{CLM} = -\frac{1}{M}\sum_{m=1}^{1}\sum_{c\epsilon C} \log P(a_c) \tag{11}$$

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

Donde C denota el subconjunto de variables, y el método degenera a pseudoverosimilitud cuando cada subconjunto es solo una variable. Si solo hay un subconjunto que contiene todas las variables, entonces la función de probabilidad compuesta es equivalente a la función de probabilidad original. Por lo tanto, la capacidad de aproximación de la función de verosimilitud compuesta se encuentra entre la pseudoverosimilitud y la verosimilitud original. El clmDCA primero calcula los parámetros de la estimación de pseudoverosimilitud como los parámetros iniciales del modelo de verosimilitud compuesto y luego actualiza los parámetros para obtener la solución final. Los resultados experimentales (Zhang, 2018) mostraron que el algoritmo clmDCA es superior al método de pseudoverosimilitud, es decir, plmDCA.

Métodos supervisados para la predicción del mapa de contactos

Los métodos de predicción basados en la coevolución presentan dos desventajas notables. En primer lugar, el método convencional de análisis de acoplamiento directo solo puede estimar la relación lineal entre las posiciones en el MSA. En segundo lugar, los algoritmos de predicción coevolutivos solo utilizan información de su propio MSA y no aprovechan información de otras fuentes. Esto puede limitar la capacidad de los algoritmos coevolutivos para predecir con precisión los mapas de contacto de residuos de proteínas, especialmente cuando se dispone de un número limitado de secuencias homólogas. Para abordar estos desafíos, los métodos supervisados basados en aprendizaje automático pueden integrar características coevolutivas y otra información relevante de secuencia y estructura en los datos de entrenamiento. Como resultado, se pueden utilizar algoritmos avanzados de clasificación para mejorar la precisión de la predicción (Jones, 2014). La Tabla 2 presenta una selección de métodos supervisados representativos basados en aprendizaje automático.

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

Tabla 2

Descripción general de los modelos representativos de aprendizaje automático supervisado para la predicción de la geometría entre residuos.

Nombre	Año	Coevolución cruda	Modelo.ML	Términos Geométricos	Disponibilidad
SVMcon	2007	No	SVM	Contacto	http://sysbio.rnet.missouri.edu/multicom_toolbox/SVMcon%201.0.html
Li et al.	2011	No	RF	Contacto	
MetaPSICOV	2014	No	ANN	Contacto	https://github.com/psipred/metapsicov
Plmconv	2016	Si	CNN	Contacto	
Nebcon	2017	No	BNN	Contacto	https://zhanggroup.org/NeBcon/
RaptorX (CASP12)	2017	No	Residual CNN	Contacto	
DNCON2	2017	No	CNN	Contacto	https://github.com/multicom-toolbox/DNCON2
DeepContact	2018	No	CNN	Contacto	https://github.com/largelymfs/deepcontact
DeepCOV	2018	Si	CNN	Contacto	https://github.com/psipred/DeepCov
ResPRE	2019	Si	Residual CNN	Contacto	https://zhanggroup.org/ResPRE
RaptorX (CASP13)	2019	No	Residual CNN	Distancia	https://github.com/j3xugit/RaptorX-Contact
AlphaFold	2020	Si	Residual CNN	Distancia	https://github.com/deepmind/deepmind-research/tree/master/alphafold_casp13
trRosetta	2020	Si	Residual CNN	Distancia, Orientacion	https://github.com/gjoni/trRosetta
TripletRes	2021	Si	Residual CNN	Contacto	https://zhanggroup.org/TripletRes
RaptorX (CASP14)	2021	Si	Residual CNN	Distancia, Orientacion	https://github.com/j3xugit/RaptorX-3DModeling
DeepPotential	2021	Si	Residual CNN	Distancia, Orientacion, H-Bond	https://zhanggroup.org/DeepPotential

Nota: elaboración propia.

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

Los algoritmos clásicos utilizados para la predicción de contactos en proteínas suelen seleccionar un par de residuos como muestra y extraer características contextuales a través de una técnica de ventana deslizante. El clasificador utilizado por estos predictores incluye bosques aleatorios (Li, 2011), máquinas de vectores de soporte (Cheng, 2007) y redes totalmente conectadas (Eickholt, 2012). Un ejemplo representativo es el método MetaPSICOV (Jones, 2014), el cual combina tres algoritmos de análisis coevolutivo como características y utiliza redes neuronales artificiales para aprender patrones de contacto de residuos. Además de las características coevolutivas, tales como PSICOV, CCMpred y mfDCA (Morcos, 2011), MetaPSICOV también incluye características estadísticas extraídas de MSA, tales como características estadísticas de uno y dos sitios. Asimismo, emplea PSIPRED (Mcguffin, 2000) y SOLVPRED para predecir la probabilidad de estructura secundaria de la secuencia y la accesibilidad solvente como características. Este método utiliza ventanas deslizantes de varios tamaños para extraer información local de cada par de residuos, y también incluye características globales, como las probabilidades de distribución de la composición de aminoácidos, la estructura secundaria promedio de proteínas y la accesibilidad solvente promedio. El método aplica una estrategia de entrenamiento de dos etapas, donde la primera etapa contiene una capa oculta de 55 nodos y genera una probabilidad de contacto residual. La segunda etapa utiliza el mismo modelo de red para la corrección de la puntuación. La precisión del método supera a otros algoritmos en CASP11, lo que demuestra la eficacia del algoritmo de aprendizaje automático supervisado (Monastyrskyy, 2016).

En CASP12, Wang et al. (Wang, 2017) propusieron RaptorX-Contact, una técnica para predecir mapas de contacto que utiliza redes neuronales convolucionales residuales (CNN), aprovechando el éxito de las redes neuronales residuales (ResNet) en la clasificación de imágenes (He, 2016). Dado que las muestras positivas y negativas de los datos de entrenamiento están muy desequilibradas, este método utiliza una función de pérdida de entropía cruzada ponderada para reducir el sesgo de categoría. El término de regularización L2 también se empleó para restringir el sobreajuste de los parámetros de la red neuronal ultra profunda. La prueba ciega en CASP12 demostró que RaptorX-Contact mejoró

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

significativamente la precisión de la predicción de los mapas de contacto de proteínas (Schaarschmidt, 2018).

En la edición CASP12, DNCON2 (Adhikari, 2017) y DeepContact (Liu, 2018) implementaron el uso de redes neuronales convolucionales (CNN) como parte fundamental de su modelo de aprendizaje profundo, y por lo tanto, fueron considerados como los métodos más destacados (Schaarschmidt, 2018). Sin embargo, los métodos convencionales emplean el procesamiento posterior del análisis coevolutivo como característica principal, lo que podría llevar a una posible pérdida de información. Después del análisis coevolutivo, se puede obtener una submatriz de 21x21 para cada par de residuos, pero el procesamiento posterior no considera los pesos ni las propiedades magnéticas de las entradas. Una manera adecuada de extraer información del análisis coevolutivo fue introducida por primera vez en plmconv (Golkov, 2016), que combinó características de acoplamiento sin procesar obtenidas por maximización de pseudoverosimilitud con una CNN. Luego, se propuso DeepCOV (Jones, 2018) para aprender patrones de contacto directamente de las matrices de covarianza sin procesar. ResPRE (Li, 2019) utiliza una estimación de cresta de la inversa de la matriz de covarianza para eliminar el ruido de transición en la matriz de covarianza. Con una sola matriz de precisión sin procesar, ResPRE superó a los métodos de alto rendimiento en CASP12 (Li, 2019). En CASP13, TripletRes (Li, 2021) fusiona un triplete de características de matriz coevolutiva sin procesar. Estas características incluyen la matriz de acoplamiento del modelo de Potts maximizado con pseudoverosimilitud, la matriz de covarianza y la estimación de la cresta de la matriz de precisión. Se utilizó una CNN residual profunda para fusionar el triplete de características, lo que permitió clasificar a TripletRes como uno de los métodos de mayor rendimiento para la predicción del mapa de contacto en CASP13 (Li, 2019).

Después de CASP13, la estrategia de utilizar la matriz de análisis coevolutivo sin procesar fue ampliamente adoptada por otros métodos representativos, por ejemplo, trRosetta (Yang, 2020), tFold (Shen, 2021), DeepPotential (Li, 2021) y otros participantes en CASP14 (Xu, 2021). Uno de los avances más significativos en CASP13 fue la introducción y predicción exitosa del concepto de predicción de distancia por parte de AlphaFold (Senior, 2020) y RaptorX (Xu, 2019). La predicción de la distancia se considera como un problema

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

de clasificación multiclase, y se pueden entrenar redes neuronales completas con pérdida de entropía cruzada. La formulación general consiste en discretizar la distancia real por debajo de un umbral determinado en un histograma de distancia, asignando una clase adicional a aquellos pares de residuos cuyas distancias están por encima del umbral. trRosetta también predice las distribuciones para el ángulo diedro ω y θ y el ángulo ϕ. Los ángulos también se discretizan en contenedores con un intervalo fijo, similar a la predicción de distancia. La inclusión de dichos términos de orientación mejoró el rendimiento de la predicción de la estructura de la proteína. En CASP14, DeepPotential, que fue una extensión de TripletRes que incluyó términos de distancia, orientación y enlace H (Yang, 2015) con aprendizaje multitarea profundo, fue muy eficaz para la predicción ab initio de la estructura proteica de objetivos de modelado libre (Zheng, 2021).

APLICACIÓN DE LA PREDICCIÓN DE INTERACCIÓN ENTRE RESIDUOS DE PROTEÍNAS

Una de las aplicaciones directas más impactantes de la predicción de la interacción entre residuos de proteínas es la predicción de la estructura de las proteínas. El enfoque inicial fue utilizar el mapa de contacto predicho para ayudar al plegamiento y clasificación de proteínas (Sadowski, 2011). DCAfold (Sułkowska, 2012) y EVfold (Marks, 2011) tomaron los mapas de contactos predichos como restricciones y los introdujeron en el paquete de software de dinámica molecular del SNC (Brunger, 1998) para encontrar la conformación proteica óptima. Dicho protocolo trajo una mejora significativa a las tuberías de plegamiento de proteínas en comparación con enfoques sin restricciones de contacto (Sułkowska, 2012). FRAGFOLD diseñó una nueva función potencial (Kosciolek, 2014) para cada par de residuos si PSICOV los predice como contactos, definida como:

$$E\begin{Bmatrix} -P, d \le d_{con} \\ -Pe^{-(d-d_{con})^2+P\frac{d-d_{con}}{d}}, d > d_{con} \end{Bmatrix} \quad (12)$$

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

Donde P es la probabilidad predicha para el par como contacto. $d_{con} = 8Å$ es el umbral que define los contactos. Ovchinnikov et al. (Ovchinnikov, 2014) alternativamente usaron restricciones de distancia sigmoidal al programa de plegado, Rosetta, en forma de

$$E = \frac{peso}{1 + \exp(d - d_{cut})} + intercept \tag{13}$$

Aquí el peso es proporcional a la fuerza de acoplamiento normalizada y d es la distancia entre los átomos de Cβ (Cα en el caso de la glicina) en la representación de átomos reducidos de los modelos de Rosetta. d_{cut} e $intercept$ son los otros dos parámetros estadísticos para diferentes modos de Rosetta. Además, la información de contacto de residuos también fue utilizada por QUARK (Mortuza, 2021) e I-TASSER (Zheng, 2021) para ayudar a predecir la estructura de proteínas para la competencia CASP (Zhang, 2018). Durante la simulación de la estructura de la proteína utilizando Monte Carlo, el término de energía para los contactos de residuos fue:

$$E = \begin{cases} -U_{ij} & d < d_{cut} \\ -\frac{1}{2}U_{ij}\left[1 - sin\left(\frac{d\left(\frac{d_{cut}D}{2}\right)}{D - d_{cut}}\pi\right)\right] & , d_{cut} \le d < D \\ \frac{1}{2}U_{ij}\left[1 + sin\left(\frac{d - \left(\frac{D+80}{2}\right)}{(80 - D)}\pi\right)\right] & , D \le d < 80Å \\ U_{ij'} & d \ge 80Å \end{cases} \tag{14}$$

Donde $d_{cut} = 8Å \; y \; D = 8Å + d_{well}$, $donde \; d_{well}$ es el ancho del pozo del primer término de la función sinusoidal y 80-D es el ancho del pozo del segundo término de la función sinusoidal.

El ancho del pozo (permanencia) es un parámetro crucial para determinar la velocidad a la que se juntan los residuos que se predice que estarán en contacto, y se ajustó en función de la longitud de las proteínas de entrenamiento. Uij son las puntuaciones de confianza para el residuo que define los límites inferior y superior de la energía. Con la integración de dicha energía de contacto en QUARK e I-TASSER, los métodos resultantes, C-QUARK y C-I-

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

TASSER, se clasificaron como dos de los mejores servidores en CASP12 y CASP13, respectivamente (Zheng, 2019). En CASP13, RaptorX (Xu, 2019) y AlphaFold (Senior, 2020) introdujeron la distancia predicha entre residuos y restricciones para la predicción de la estructura de proteínas. Ambos modelos predicen intervalos de distancia discretos y RaptorX estima la media y la desviación estándar a partir de las predicciones del histograma de distancia. Mientras tanto, AlphaFold construye un potencial diferenciable para un par de residuos dado mediante la interpolación de un spline cúbico a partir de la probabilidad logarítmica negativa de la distribución de distancia. Tanto RaptorX como AlphaFold lograron resultados prometedores en CASP13, y AlphaFold fue particularmente exitoso (Abriata, 2019) en la predicción de la estructura más precisa para los objetivos clasificados como los más difíciles por los organizadores de la competencia. Después de CASP13, otro método de predicción de la estructura de proteínas basado en el aprendizaje profundo, trRosetta (Yang, 2020), convirtió las distribuciones de distancia predichas en energía potencial siguiendo la idea de Dfire (Zhou, 2002). Se consideró como estado de referencia la probabilidad del último bin y la energía a distancia fue:

$$E(i) = -\ln(p_i) + \ln\left(\left(\frac{d_i}{d_N}\right)^{\alpha} P_N\right), i = 1,2, \ldots\ldots, N \qquad (15)$$

Donde pi es la probabilidad para el intervalo de distancia i-ésima. N es el número total de contenedores, α es una constante (= 1,57) para la normalización basada en la distancia y di es la distancia para el i-ésimo contenedor de distancia. Para términos de orientación, es decir, ángulos y ángulos diédricos, la energía es similar, pero sin normalización:

$$E(i) = -\ln(p_i) + \ln(P_N), i = 1,2, \ldots, N \qquad (16)$$

El suavizado de los potenciales se logra a través del uso del spline cúbico y su minimización se puede llevar a cabo mediante la aplicación de técnicas como el descenso de gradiente, incluyendo el algoritmo L-BFGS (Chaudhury, 2010). La distinción entre la predicción de la interacción de los residuos de proteínas y la predicción de la estructura de proteínas se ha vuelto difusa debido a la propuesta de varios métodos de predicción de la estructura de proteínas de extremo a extremo. Estos métodos pueden generar directamente las coordenadas de la proteína. El modelo representativo más destacado en la predicción de la estructura de proteínas en CASP14 fue AlphaFold2 (Jumper, 2021; Kinch, 2021)

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

En AlphaFold2, la estructura de cada residuo se representa mediante una matriz de rotación global y un vector de traducción en el módulo de estructura de AlphaFold2. La cadena lateral de cada residuo se puede recuperar mediante la predicción posicional de los ángulos Chi. AlphaFold2 actualiza iterativamente las representaciones de residuos mediante una operación de atención consciente de la geometría, atención de puntos invariantes (IPA). Este módulo se basa en la atención y toma información geométrica en la conformación anterior en los mapas de atención. Las salidas del módulo IPA para cada residuo son las actualizaciones de la matriz de rotación (en cuaterniones) y el vector de traducción, que se podrían aplicar a las anteriores, respectivamente.

Una vez que se obtiene la estructura, AlphaFold2 emplea una nueva función de pérdida llamada error de punto alineado con el marco (FAPE). Esta función mide los errores entre las posiciones predichas de los átomos y las coordenadas reales del terreno bajo diferentes alineaciones. Las alineaciones se definen a partir de los fotogramas predichos, es decir, la matriz de rotación y el vector de traducción, además de los fotogramas adicionales de las cadenas laterales. La función de pérdida es invariante a transformaciones rígidas, lo que facilita la optimización de la red neuronal.

AlphaFold2 también utiliza el mecanismo de atención de última generación como red troncal del módulo Evoformer en AlphaFold2. En comparación con las CNN, que solo pueden aprender información local, el mecanismo de atención tiene la capacidad de considerar la información de todas las demás variables con la codificación posicional adecuada. El módulo Evoformer también permite que las redes neuronales de transformadores basadas en la atención atiendan arbitrariamente el MSA completo en lugar de utilizar las funciones de análisis coevolutivo (en bruto). En comparación con los enfoques convencionales, el nuevo algoritmo se enfoca en secuencias más relevantes y extrae información más rica del MSA.

Además, se han empleado algunos esfuerzos de ingeniería, como el reciclado del proceso de capacitación, y el modelo resultante ha logrado un rendimiento de vanguardia en CASP14 (Jumper, 2021).

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

DISCUSIÓN Y RESULTADOS

Las interacciones entre residuos de proteínas, tales como los contactos, son una representación simplificada de la estructura de la proteína. La predicción precisa de estas interacciones es de gran importancia para la predicción de la estructura de las proteínas. Este capítulo presenta una revisión del desarrollo de algoritmos para la predicción de la interacción entre residuos, así como los últimos resultados de investigación en este campo.

El progreso rápido en la predicción de la interacción entre residuos, especialmente en los métodos basados en el aprendizaje profundo, ha llevado a la solución final del problema de predicción de la estructura de las proteínas. En este campo, se han logrado cinco hitos importantes:

1. Introducción del análisis global de coevolución, donde los modelos globales, es decir, los métodos DCA, han proporcionado un modelo probabilístico sistemático para los datos de MSA utilizando el modelo de Potts. Con una aproximación adecuada, los modelos DCA pueden eliminar de manera eficiente el ruido de transición en los métodos de análisis coevolutivos locales.
2. Aprendizaje convolucional profundo de las predicciones de contacto, donde la predicción de contactos de residuos de proteínas se formaliza con redes neuronales convolucionales, que han demostrado una ventaja significativa sobre los métodos clásicos. La introducción de redes neuronales residuales mejora aún más el rendimiento.
3. Uso de características coevolutivas en bruto, donde alimentar directamente el modelo de aprendizaje profundo con correlaciones coevolutivas sin procesar puede evitar la pérdida de información. El uso de funciones sin procesar sin el posprocesamiento corrige los problemas de los procedimientos de extracción de funciones en modelos anteriores.
4. Consideración de términos geométricos más ricos, donde los términos de geometría adicionales pueden proporcionar más información para ayudar a plegar mejor las estructuras de proteínas. Además, el aprendizaje multitarea también puede contribuir a cada una de las tareas.

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

5. Un modelo de aprendizaje efectivo de extremo a extremo, donde el modelo AlphaFold2 se considera casi resolvió el problema de predicción de la estructura de proteínas. El aprendizaje de extremo a extremo puede alimentar directamente el error de estructura al modelo. El poderoso mecanismo de atención también contribuye al éxito de AlphaFold2.

A pesar de estos logros, todavía hay un par de problemas que deben abordarse. En primer lugar, la predicción de las interacciones entre residuos entre dominios y cadenas de proteínas multidominio y complejos proteicos necesita una solución específica. En segundo lugar, se necesita más investigación para proporcionar información interpretable durante la predicción del plegamiento de proteínas basada en el aprendizaje profundo. Esperamos ver que el objetivo final de la predicción de la estructura de la proteína, es decir, la predicción directa de estructuras de proteína de alta precisión solo con secuencias de aminoácidos, se resuelva en un futuro próximo.

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

REFERENCIAS

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H. y Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, *181*(4610), 662-666.

Wüthrich, K. (2001). The way to NMR structures of proteins. *Nature Structural & Molecular Biology*, *8*(11), 923.

Taylor, K. A. y Glaeser, R. M. (1974). Electron diffraction of frozen, hydrated protein crystals. *Science*, *186*(4168), 1036-1037.

Baker, D. y Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, *294*(5540), 93-96.

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

Webb, B. y Sali, A. (2016). Comparative protein structure modeling using MODELLER. *Current Protocols in Bioinformatics*, *54*(1), 1-5.

Schwede, T., Kopp, J., Guex, N. y Peitsch, M. C. (2003). SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Research*, *31*(13), 3381-3385.

Bowie, J. U., Lüthy, R. y Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, *253*(5016), 164-170.

Jones, D. y Thornton, J. (1993). Protein fold recognition. *Journal of ComputerAided Molecular Design*, *7*(4), 439-456.

Cocco, S., Feinaur, C., Figliuzzi, M., Monasson, R. y Weigt, M. (2018). Inverse statistical physics of protein sequences: A key issues review. *Reports on Progress in Physics Physical Society*, *81*(3), 032601.

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. y Aurell, E. (2013). Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E*, 87(1), 012707.

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

Jones, D. T., Buchan, D., Cozzetto, D. y Pontil, M. (2011). PSICOV: Precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, *28*(2), 184-190.

Martin, L., Gloor, G. B., Dunn, S. D. y Wahl, L. M. (2005). Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, *21*(22), 4116-4124.

Morcos, F., Pagnani, A., Lunt, B. y Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, *108*(49), 1293-1301.

Seemayer, S., Gruber, M. y Söding, J. (2014). CCMpred - fast and precise prediction of protein residue–residue contacts from correlated mutations. *Bioinformatics*, *30*(21), 3128-3130.

Cheng, J. y Baldi, P. (2007). Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, *8*(1), 1-9.

Jones, D.T., Singh, T., Koscioleck, T. y Tetchner, S. (2014). MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long-range hydrogen bonding in proteins. *Bioinformatics*, *31*(7), 999-1006.

Schneider, M. y Brock, O. (2014). Combining physicochemical and evolutionary information for protein contact prediction. *Plos One*, *9*(10), e108438.

Adhikari, B., Hou, J. y Cheng, J. (2017). DNCON2: Improved protein contact prediction using twolevel deep convolutional neural networks. *Bioinformatics*, *34*(9),1466-1472.

He, B., Mortuza, S. M., Wang, Y., Shen, H. B. y Zhang, Y. (2017). NeBcon: Protein contact map prediction using neural network training coupled with naïve Bayes classifiers. *Bioinformatics*, *33*(15), 2296.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Zidek, A., Nelson, A., Bridglan, A., Penedones, H., Peterson, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D., Silver, D., Kavukcuoglu, K. y Hassabis, D. (2020). Improved

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

protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706-710.

Xu, J. (2019). Distance-based protein folding powered by deep learning. *Proceedings of the National Academy of Sciences*, *116*(34), 16856-16865.

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S. y Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, *117*(3), 1496-1503.

Li, Y., Zhang, C., Zheng, W., Zhou, X., Bell, E. W., Yu, B. J. y Zhang, Y. (2021). Protein inter-residue contact and distance prediction by coupling complementary coevolution features with deep residual networks in CASP14. *Proteins: Structure, Function, and Bioinformatics*, *89*(12), 1911-1921.

Moult, J. A. (2005). Decade of CASP: Progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology*, *15*(3), 285-289.

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

Yanofsky, C., Horn, V. y Thorpe, D. (1964). Protein structure relationships revealed by mutational analysis. *Science*, *146*(3651), 1593-1594.

Korber, B. T., Farber, R. M., Wolpert, D. H. y Lapedes, A. S. (1993). Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: An information theoretic analysis. *Proceedings of the National Academy of Sciences*, *90*(15), 7176-7180.

Kass, I. y Horovitz, A. (2002). Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins: Structure, Function, and Bioinformatics*, *48*(4), 611-617.

Lapedes, A. S., Giraud, B. G., Liu, L. C. y Stormo, G. D. (1999). Correlated mutations in models of protein sequences: Phylogenetic and structural effects. *Lecture Notes-Monograph Series*, *33*, 236-256.

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

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*, *106*(1), 67.

Li, Y., Hu, J., Zhang, C., Yu, D. J. y Zhang, Y. (2019). ResPRE: High-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics*, *35*(22), 4647-4655.

Kamisetty, H., Ovchinnikov, S. y Baker, D. (2013). Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(39), 15674-15679.

Zhang, H., Zhang, Q., Ju, F., Zhu, J., Gao, Y., Xie, Z., Deng, M., Sun, S., Zheng, W. M. y Bu, D. (2018). Predicting protein inter-residue contacts using composite likelihood maximization and deep learning. *BMC Bioinformatics*, *20*, 537.

Liu, D. C. y Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, *45*(1-3), 503-528.

Hestenes, M. R. y Stiefel, E. (1952). Methods of conjugate gradients for solving linear systems. *NBS Washington*, *49*, 409-435.

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

Golkov, V., Skwark, M. J., Golkov, A., Dosovitskiy, A., Brox, T., Meiler, J. y Cremers, D. (2016). Protein contact prediction from amino acid co-evolution using convolutional networks for graph-valued images. *Advances in Neural Information Processing Systems*, *29*, 4222-4230.

Wang, S., Sun, S., Li, Z., Zhang, R. y Xu, J. (2017). Accurate De Novo prediction of protein contact map by ultradeep learning model. *Plos Computational Biology*, *13*(1), e1005324.

Liu, Y., Palmedo, P., Ye, Q., Berger, B. y Peng, J. (2018). Enhancing evolutionary couplings with deep convolutional neural networks. *Cell Systems*, *6*(1), 65.

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

Jones, D. T. y Kandathil, S. M. (2018). High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics*, *34*(19), 3308-3315.

Li, Y., Zhang, C., Bell, E. W., Zheng, W., Zhou, X., Yu, D. J. y Zhang, Y. (2021). Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. *Plus Computational Biology*, *17*(3), e1008865.

Xu, J., Mcpartlon, M. y Li, J. (2021). Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nature Machine Intelligence*, *3*, 601-609.

Eickholt, J. y Cheng, J. (2012). Predicting protein residue–residue contacts using deep networks and boosting. *Bioinformatics*, *28*(23), 3066-3072.

Mcguffin, L. J., Bryson, K. y Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics*, *16*(4), 404-405.

Monastyrskyy, B., D'Andrea, D., Fidelis, K., Tramontano, A. y Kryshtafovych, A. (2016). New encouraging developments in contact prediction: Assessment of the CASP11 results. *Proteins: Structure, Function, and Bioinformatics*, *84*(s1), 131-144.

He, K., Zhang, X., Ren, S. y Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, IEEE,

Schaarschmidt, J., Monastyrskyy, B., Kryshtafovych, A. y Bonvin, A. M. J. J. (2018). Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. *Proteins: Structure, Function, and Bioinformatics*, *86*, 51-66.

Li, Y., Zhang, C., Bell, E. W., Yu, D. J. y Zhang, Y. (2019). Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. *Proteins: Structure, Function, and Bioinformatics*, *87*(12), 1082-1091.

Shen, T., Wu, J., Lan, H., Zheng, L., Pei, J., Wang, S., Liu, W. y Huang, J. (2021). When homologous sequences meet structural decoys: Accurate contact prediction by tFold

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

in CASP14 — (tFold for CASP14 contact prediction). *Proteins: Structure, Function, and Bioinformatics*, *89*(12), 1901-1910.

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.

Zheng, W., Li, Y., Zhang, C., Zhou, X., Pearce, R., Bell, E. W., Huang, X. y Zhang, Y. (2021). Protein structure prediction using deep learning distance and hydrogen-bonding restraints in CASP14. *Proteins: Structure, Function, and Bioinformatics*, *89*(12), 1734-1751.

Sadowski, M. I., Maksimiak, K. y Taylor, W. R. (2011). Direct correlation analysis improves fold recognition. *Computational Biology and Chemistry*, *35*(5), 323-332.

Taylor, W. R., Jones, D. T. y Sadowski, M. I. (2012). Protein topology from predicted residue contacts. *Protein Science*, *21*(2), 299-305.

Sułkowska, J. I., Morcos, F. y Weigt, M. (2012). Genomics-aided structure prediction. *Proceedings of the National Academy of Sciences*, *109*(26), 10340.

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R. y Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. *PloS One*, *6*(12), e28766.

Brunger, A. T., Adams, P. D., Clore, G. M., Delano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. y Warren, G. L. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallographica Section D: Biological Crystallography*, *54*(5), 905-921.

Kosciolek, T. y Jones, D. T. (2014). De novo structure prediction of globular proteins aided by sequence variation-derived contacts. *PloS One*, *9*(3), e92197.

Ovchinnikov, S., Kamisetty, H. y Baker, D. (2014). Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *eLife*, *3*, e02030.

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

Mortuza, S., Zheng, W., Zhang, C., Li, y Pearce, R. (2021). Improving fragment-based ab initio protein structure assembly using low-accuracy contact-map predictions. *Nature Communications*, *12*(1), 1-12.

Zheng, W., Zhang, C., Li, Y., Pearce, R. y Bell, E. W. (2021). Folding non-homologous proteins by coupling deep-learning contact map with I-TASSER assembly simulations. *Cell Reports Methods*, *1*, 100014.

Zhang, C., Mortuza, S. M., He, B., Wang, Y. y Zhang, Y. (2018). Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. *Proteins: Structure, Function, and Bioinformatics*, *86*, 136-151.

Zheng, W., Li, Y., Zhang, C., Pearce, R., Mortuza, S. M. y Zhang, Y. (2019). Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins: Structure, Function, and Bioinformatics*, *87*(12), 1149-1164.

Abriata, L. A., Tamó, G. E. y Peraro, M. (2019). A further leap of improvement in tertiary structure prediction in CASP13 prompts new routes for future assessments. *Proteins: Structure, Function, and Bioinformatics*, *87*(12), 1100-1112.

Zhou, H. y Zhou, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science*, *11*(11), 2714-2726.

Chaudhury, S., Lyskov, S. y Gray, J. J. (2010). PyRosetta: A script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics*, *26*(5), 689-691.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O. y Tunyasuvunakool, K. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583-589.

Kinch, L. N., Pei, J., Kryshtafovych, A., Schaeffer, R. D. y Grishin, N. V. (2021). Topology evaluation of models for difficult targets in the 14th round of the critical assessment of protein structure prediction (CASP14). *Proteins: Structure, Function, and Bioinformatics*, *89*(12), 1673-1686.

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