Plegamiento de proteínas 2D.

Inteligencia Artificial – Ingeniería del Software

Curso 2021/2022

Propuesta de trabajo

Ana Belén Romero-Losada.

1. Introducción y objetivos.

Las proteínas son unas de las macromoléculas más importantes en la biología molecular. Están presentes en todas las formas de vida conocidas (e incluso en los virus) y son las encargadas de llevar a cabo casi todos los procesos necesarios para que la vida siga adelante.

Las proteínas están formadas por cadenas de aminoácidos que se doblan y acoplan entre sí espontáneamente, en un proceso llamado plegamiento de proteínas, para formar una estructura tridimensional relacionada con la función biológica de la proteína.

Uno de los desafíos de la biología molecular, conocido como el "**problema del plegamiento de las proteínas"** consiste en entender cómo la secuencia de aminoácidos determina la estructura tridimensional. Para solucionar este problema, es necesario comprender la termodinámica de las fuerzas interatómicas que resultan en una estructura estable y el mecanismo por el que las proteínas alcanzan su configuración final con extrema rapidez.

Las estructuras de las proteínas se determinan habitualmente de forma experimental mediante técnicas que son costosas y pueden requerir mucho tiempo. Durante los últimos sesenta años solo se han identificado las estructuras de unas 170 000 proteínas, de las más de doscientos millones que se calcula que existen en todas las formas de vida conocidas. El poder predecir la estructura de las proteínas sin más información que la secuencia de aminoácidos sería de gran ayuda para avanzar en la investigación científica. Sin embargo, la **paradoja de Levinthal** muestra que, si bien una proteína se puede plegar en milisegundos, el tiempo que lleva calcular todas las estructuras posibles al azar para determinar la estructura más óptima es más largo que la edad del universo conocido. En resumen, se trata de un **problema de optimización** en el que hay innumerables soluciones con una puntuación distinta cada una.

A lo largo de los años, los investigadores han aplicado numerosos métodos computacionales al problema de la predicción de la estructura de las proteínas, pero la precisión de los modelos generados jamás se había acercado a la de las estructuras determinadas por técnicas experimentales, excepto para proteínas pequeñas y simples. Hasta la entrada en escena de **AlphaFold2** (2020), un programa de inteligencia artificial (IA) desarrollado por DeepMind de Alphabets/Google que realiza predicciones de la estructura de las proteínas mediante Deep Learning. Este hito multidisciplinar ha marcado un antes y un después en la comunidad científica.

En este trabajo se propone implementar uno de los primeros modelos de plegamiento de proteínas, menos complejos que AlphaFold2, pero que construyeron la base para abordar el histórico problema del plegamiento de las proteínas desde la computación y la inteligencia artificial.

Para ello, se buscará alcanzar los siguientes **objetivos** específicos:

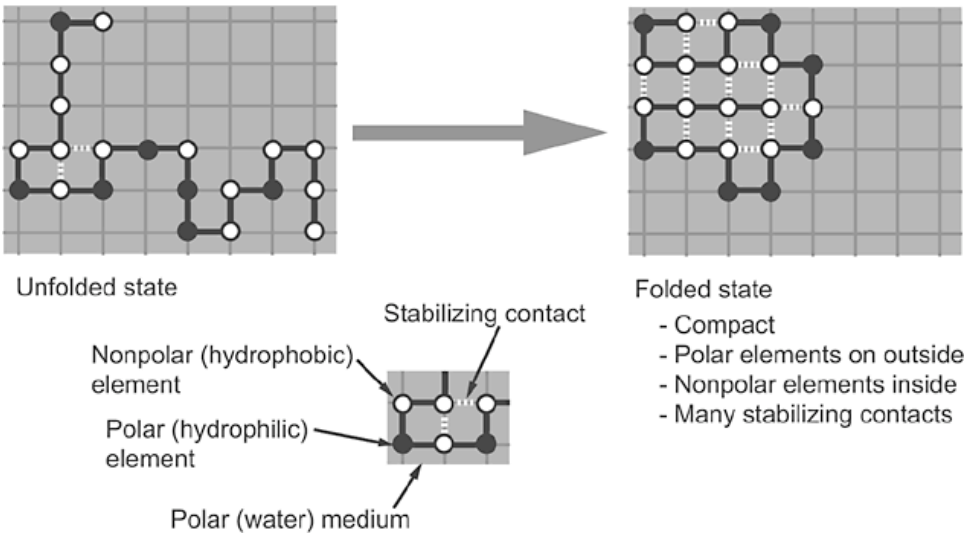
- Comprender la teoría básica detrás del plegamiento de proteínas y la paradoja de Levinthal. Comparar estos conceptos con un problema de optimización.
- Identificar las distintas disciplinas necesarias para estudiar la paradoja de Levinthal. Entender la importancia de la multidisciplinariedad.
- Aprender a implementar un algoritmo en base a un modelo de plegamiento de proteínas 2D.
- Realizar gráficos con *pyplot* para representar los plegamientos de proteínas 2D.
- Aprender a usar la base de datos de proteínas *UNIPROT*.
- Documentar el trabajo realizado usando un formato de artículo científico.
- Realizar una presentación de los resultados obtenidos.

2. Descripción del trabajo.

2.1. Metodología.

Para el comprender la teoría básica de este proceso biológico se usará el libro *Bioquímica* de Stryer (disponible online en distintas plataformas y en físico en CRAI), concretamente el capítulo 2 titulado *Composición y estructura de las proteínas*.

Además, se pueden usar los artículos Shakhnovich & Gutin 1993, Dill et al. 1995, Moreno-Hernández & Levitt 2012 donde se describen los primeros modelos de plegamientos de proteínas 2D. Usaron un alfabeto reducido en donde los aminoácidos son polares/hidrofílicos (con tendencia a mezclarse con el agua) o apolares/hidrofóbicos (repelen el agua) y donde las cadenas se pliegan en una malla bidimensional. En general se acepta que estos modelos reproducen las características más importantes del proceso de plegamiento real con la ventaja de ser más manejables.



La solubilidad de los aminoácidos se mide calculando la variación de Energía Libre en contacto con agua a pH = 7. Esto va a determinar si el aminoácido es hidrofílico o hidrofóbico (Energía libre mayor que –1.5). A

continuación, se facilita un diccionario con los 20 aminoácidos y sus correspondientes valores de la variación de Energía libre:

```
aa_deltaG = { 'A': 1,   # Alanine
              'C': 0.17, # Cysteine
              'D': -3,   # Aspartic Acid
              'E': -2.6, # Glutamic Acid
              'F': 2.5,  # Phenylalanine
              'G': 0.67, # Glycine
              'H': -1.7, # Histidine
              'I': 3.1,  # Isoleucine
              'K': -4.6, # Lysine
              'L': 2.2,  # Leucine
              'M': 1.1,  # Methionine
              'N': -2.7, # Asparagine
              'P': -0.29,# Proline
              'Q': -2.9, # Glutamine
              'R': -7.5, # Arginine
              'S': -1.1, # Serine
              'T': -0.75,# Threonine
              'V': 2.3,  # Valine
              'W': 1.5,  # Tryptophan
             'Y': 0.08  # Tyrosine}
```

En esta práctica vamos a implementar un sencillo algoritmo de plegamiento de proteínas en 2D en base a la hidrofobicidad de los aminoácidos. Para ello, se pide implementar un **algoritmo de Enfriamiento Simulado** (Simulated annealing) para resolver este problema de optimización. Este algoritmo, como otros muchos, se trata de un método de optimización inspirado en un proceso natural, es decir, fue creado basándose en la observación de una excelente optimización de un fenómeno que tiene lugar en la naturaleza. Concretamente, el algoritmo de enfriamiento simulado está inspirado en el proceso físico por el que las moléculas de un metal fundido se reordenan al enfriarse de forma que generan una estructura sólida y compleja con el menor estado energético posible. Se recomiendan los libros *Algorithms for Optimization* de Kochenderfer (disponible en físico en la biblioteca de Informática) y *Adaption of simulated annealing to chemical de* Kalivas (disponible online en fama.us) como apoyo para conocer en profundidad las bases de los algoritmos de enfriamiento simulado.

Con estos conceptos como base, se propone resolver los siguientes problemas:

Problema 1.

- Se pide implementar una función **get_spatial_dic(protein, structure)** que recibe una cadena representando una proteína (letras de aminoácidos) y otra cadena representando su estructura (I para el aminoácido inicial, N, S, E o W según la posición relativa de un aminoácido con respecto al anterior) y devuelve un diccionario o un diccionario vacío si existen solapamientos. Las claves de dicho diccionario serán tuplas de dos números enteros representando coordenadas espaciales y los valores serán letras de aminoácidos.

Problema 2.

- Se pide implementar una función **is_hydrophobic(aa).**
- Se pide implementar una función **get_score(dic)** que reciba un diccionario representando la estructura espacial de una proteína y devuelva su puntuación. La puntuación de un aminoácido

será ΔG * N (si el aminoácido no es hidrofóbico) y ΔG * N + 10 * N (si el aminoácido es hidrofóbico). Siendo N el número de posiciones adyacentes libres.

Problema 3.

- Se pide implementar una función **fold(structure, pos, angle)** que recibe una estructura, una posición de plegado (desde donde se comienza a plegar) y el ángulo que puede ser 90 o –90.
- Se pide implementar una función **get_successors(protein,structure)** que, dada una proteína y su estructura, devuelva un diccionario cuyas claves son todas las posibles estructuras **válidas** tras aplicar todos los posibles plegamientos y cuyos valores sean los correspondientes diccionarios espaciales obtenidos con **get_spatial_dic**.

Problema 4.

- Usando las funciones creadas en los problemas anteriores, implementa el **algoritmo de Enfriamiento Simulado** (Simmulating Annealing) para resolver el problema del plegado 2D como un problema de optimización.

Problema 5.

Dada una proteína podemos predecir un plegamiento 2D de la misma gracias a los algoritmos implementados durante el desarrollo de este trabajo.

- Se pide probar estos algoritmos con al menos otras tres proteínas sencillas cuya secuencia de aminoácidos y descripción sea extraída desde *UNIPROT*.
- Se pide hacer una representación gráfica de las estructuras 2D resultante con *pyplot.* Coloreando de distinto color los aminoácidos dependiendo de si son hidrofóbicos o no.
- Por último, OPCIONALMENTE para optar a nota extra, se pide intentar implementar un gradiente de color dependiendo del valor ΔG del correspondiente aminoácido.

2.1. Documentación y entrega.

El trabajo deberá documentarse siguiendo un formato de artículo científico, con una extensión mínima de 6 páginas. En la página web de la asignatura se pueden encontrar plantillas donde se sugiere una estructura general. Estas plantillas siguen el formato de los IEEE conference proceedings, cuyo sitio web guía para autores ofrece información más detallada. El documento entregado deberá estar en formato PDF.

En el caso concreto de este trabajo, la memoria deberá al menos incluir: introducción; funcionamiento de los algoritmos de plegamiento de proteínas 2D y como contribuyeron a los algoritmos actuales (AlphaFold2); explicación y gráficas de cada uno de los plegamientos obtenidos, haciendo énfasis en la multidisciplinariedad de esta metodología; conclusiones; bibliografía (al menos dos elementos que no aparezcan como bibliografía en este documento). En ningún caso debe incluirse código en la memoria. La entrega del trabajo consistirá de la memoria del trabajo y el código implementado (cuadernos de Jupyter). Ambos deben subirse a la página de la asignatura en un único fichero comprimido zip.

3. Evaluación del trabajo.

Para que el trabajo pueda ser evaluado se debe haber implementado correctamente el algoritmo propuesto y haberlo usado para predecir la estructura 2D de al menos 3 proteínas distintas. Para la evaluación del trabajo se tendrán en cuenta los siguientes criterios, considerando una nota total máxima de 4 puntos:

Memoria del trabajo (hasta 1.5 puntos): se valorará la claridad de las explicaciones, el razonamiento de las decisiones, el análisis y presentación de resultados y el correcto uso del lenguaje. La elaboración de la memoria debe ser original, por lo que no se evaluará el trabajo si se detecta cualquier copia del contenido.

Código fuente (hasta 1.5 puntos): se valorará la claridad y buen estilo de programación, corrección y eficiencia de la implementación y calidad de los comentarios. El código debe ser original, por lo que no se evaluará el trabajo si se detecta código copiado o descargado de internet.

Presentación y defensa (hasta 1 puntos): se valorará la claridad de la presentación y la buena explicación de los contenidos del trabajo, así como, las respuestas a las preguntas realizadas por la profesora.

Apartado opcional (hasta 0,5 puntos extra): se valorará la resolución del apartado opcional para subir la nota. En ningún momento se podrá superar la nota de 4 puntos en la evaluación completa.

IMPORTANTE: Cualquier plagio, compartición de código o uso de material que no sea original y del que no se cite convenientemente la fuente, significará automáticamente la calificación de cero en la asignatura para todos los alumnos involucrados. Por tanto, a estos alumnos no se les conserva, ni para la actual ni para futuras convocatorias, ninguna nota que hubiesen obtenido hasta el momento. Todo ello sin perjuicio de las correspondientes medidas disciplinarias que se pudieran tomar.

4. Bibliografía.

Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* (2021). https://doi.org/10.1038/s41586-021-03819-2

Shakhnovich & Gutin *et al*. Engineering of stable and fast-folding sequences of model proteins. *PNAS* (1993). https://doi.org/10.1073/pnas.90.15.7195

Dill *et al*. Principles of protein folding: A perspective from simple exact models. *Protein Science* (1995) https://doi.org/10.1002/pro.5560040401

Moreno-Hernandez & Levitt. Comparative modeling and protein-like features of hydrophobic-polar models on a two-dimensional lattice. *Proteins*. (2012) https://doi.org/10.1002/prot.24067

Berg, J. M., Tymoczko, J. L., Stryer, L., & Stryer, L. Biochemistry. *New York: W.H. Freeman*. (2002).

Kapoor, A. Hands-On Artificial Intelligence for IoT : Expert Machine Learning and Deep Learning Techniques for Developing Smarter IoT Systems, *Packt Publishing*, Limited, (2019). https://ebookcentral.proquest.com/lib/uses/detail.action?docID=5675583.

Kochenderfer, & Wheeler, T. A. Algorithms for optimization. *The Mit Press*. (2019) ISBN 9780262039420.

KALIVAS, J.H. Adaption of simulated annealing to chemical optimization problems. *New York: Elsevier.* (1995) ISBN 1-281-05813-0.