UNIVERSIDAD DE LA HABANA

Facultad de Matemática y Computación

PROYECTO FINAL
DISEÑO Y ANÁLISIS DE ALGORITMOS

Diseño Computacional de Proteínas

Autor: Lidier Robaina Caraballo

Grupo: C-411

10 de febrero de 2025

Índice general

Capítulo 1

Introducción

La bioinformática emerge como un campo revolucionario en la intersección entre la biología, la química y la ciencia de la computación, ofreciendo herramientas innovadoras para abordar desafíos científicos y tecnológicos de alto impacto. Las proteínas, como máquinas moleculares esenciales para la vida, desempeñan roles críticos en procesos biológicos, aplicaciones médicas (como el desarrollo de fármacos y terapias) y soluciones industriales (como biocatalizadores y materiales biodegradables). Sin embargo, el diseño tradicional de proteínas, basado en métodos experimentales de ensayo y error, resulta costoso, lento y limitado en complejidad. Aquí es donde los algoritmos computacionales se convierten en un aliado indispensable: permiten modelar, predecir y optimizar estructuras proteicas con precisión, acelerando el descubrimiento de soluciones biológicas personalizadas.

La **motivación** de este proyecto radica en dos pilares fundamentales. Primero, la necesidad de explorar estrategias algorítmicas eficientes para resolver problemas NP-duros asociados al diseño de proteínas, como el plegamiento tridimensional y la estabilidad termodinámica. Segundo, la oportunidad de contribuir al avance de áreas como la medicina personalizada, la bioingeniería y la sostenibilidad ambiental, donde proteínas diseñadas a la medida podrían transformar paradigmas actuales.

Los **objetivos** del proyecto se centran en:

1. Modelar el problema de diseño de proteínas desde una perspectiva algorítmica, identy pueden realizar ificando sus componentes críticos (espacio de búsqueda, funciones de energía, restricciones biológicas).

2. Implementar y comparar algoritmos que aborden el diseño de secuencias, evaluando su eficiencia (tiempo y espacio) y efectividad (precisión y estabilidad de las soluciones).

3. Analizar limitaciones y ventajas de enfoques clásicos frente a métodos modernos basados en aprendizaje automático.

Capítulo 2

Definición del Problema

2.1 Marco Teórico

Las proteínas son macromoléculas esenciales para la vida dado que están involucradas en casi todas las funciones estructurales, catalíticas, sensoriales y regulatorias de los organismos. Están formadas por una secuencia de compuestos orgánicos llamadas aminoácidos, y su función está determinada fundamentalmente por la estructura tridimensional que adopta la cadena de aminoácidos.

El objetivo del diseño de proteínas es encontrar, dentro de una colección de proteínas, la que más probabilidades tiene de ajustarse a una función. Puesto que existen veinte aminoácidos posibles para cada posición en una secuencia proteica, y cada uno de ellos admite diversas variantes estructurales, la cantidad de combinaciones factibles a evaluar supera las posibilidades de cualquier método experimental, incluso en el caso de secuencias muy cortas.

2.1.1. Esqueleto

Como ilustra la figura 2.1, los aminoácidos están formado por un núcleo común (átomos de H, C, N y O) y una cadena lateral (R) que es específica para cada uno de los 20. En una proteína, los núcleos de los aminoácidos están enlazados en una secuencia que forma el *esqueleto* de la proteína.

El esqueleto es relativamente rígido y define la estructura tridimensional de la proteína (figura 2.2), por tanto se asume que la proteína resultante del diseño conservará el plegamiento global de la estructura base elegida. Es decir, se considera que el esqueleto de la proteína está fijo y es entrada del problema del diseño de proteínas.

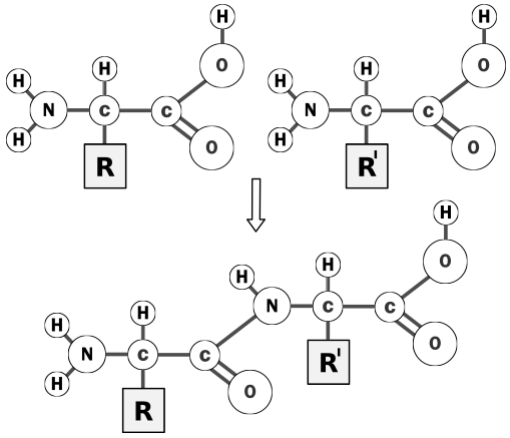

Figura 2.1: Representación de cómo dos aminoácidos, con cadenas laterales específicas R y R', enlazan sus núcleos para formar una cadena

(a) Modelo con cadenas laterales

(b) Esqueleto

Figura 2.2: Modelo 3D de una proteína

2.1.2. Rotámeros

Debido a la rotación alrededor de los enlaces simples, las cadenas laterales de los aminoácidos pueden adoptar distintas conformaciones tridimensionales, que reciben el nombre de *rotámeros*. Hay una cantidad infinita de rotámeros posibles, puesto que las moléculas pueden girar alrededor del eje de forma continua. No obstante, suele ser suficiente considerar un conjunto discreto y representativo de rotámeros (figura 2.3), el cual se puede determinar a través de un análisis estadístico de las conformaciones reales presentes en las bases de datos de estructuras de proteínas.

En el problema del diseño de proteínas, los rotámeros constituyen el espacio de búsqueda primario cuando el esqueleto está fijo.

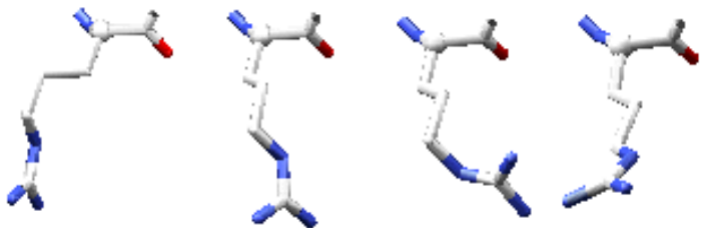

Figura 2.3: Rotámeros del aminoácido arginina: algunas posibles geometrías de la cadena lateral

2.1.3. Funciones de energía

La estabilidad y la eficiencia funcional de una proteína están correlacionadas con su energía, por lo tanto, el objetivo es encontrar la conformación que posea la energía total mínima.

La energía total de la proteína está dada por la energía del esqueleto, la energía de interacción entre los rotámeros y el esqueleto, y la energía de interacción entre los diferentes rotámeros. Cuando el esqueleto está fijo, la minimización de la energía depende únicamente de la energía de interacción entre los rotámeros y entre estos y el esqueleto. Sean i_r el rotámero en la posición i, $E(i_r)$ la energía de interacción entre el rotámero i y el esqueleto, y $E(i_r, j_{r'})$ la energía de interacción entre los rotámeros r en la posición i y r' en la posición j, la fórmula a minimizar se convierte en

$$E = \sum_i E(i_r) + \sum_i \sum_{j<i} E(i_r, j_{r'})$$

2.2 Modelación como problema de teoría de grafos

Una proteína con k residuos puede ser representada por un grafo k-partito $G = (V, E)$, de forma tal que:

- hay una partición V_i por cada residuo i
- hay un vértice $v \in V_i$ por cada rotámero candidato del residuo i
- la arista $\langle u, v \rangle$ denota la interacción entre los rotámeros u y v
- el vértice v tiene un costo c_v que es igual a la energía de interacción entre el rotámero v y el esqueleto
- la arista $\langle u, v \rangle$ tiene un costo $c_{\langle u,v \rangle}$ que es igual a la energía de interacción entre el rotámero u y el rotámero v

El problema del **Diseño Computacional de Proteínas (DCP)** se define como:

> Dado un grafo k-partito $G = (V, E)$, $V = V_1 \cup V_2 \cup ... \cup V_k$, ponderado en vértices y aristas, hallar la asignación $a : [k] \to V$ con $a(i) \in V_i$, tal que el costo
>
> $$\sum_i c_{a(i)} + \sum_i \sum_{j<i} c_{\langle a(i),a(j)\rangle}$$
>
> del grafo inducido por el conjunto de vértices $\{a(i)\}$ sea mínimo.

Capítulo 3

Análisis de complejidad

Para demostrar que **DCP** es un problema **NP-Hard**, se realiza una reducción desde el **problema de Clique**:

3.1 DCP $\in$ NP-Hard

3.1.1. Reducción Clique $\propto$ DCP

Sea $G = (V, E)$ un grafo no dirigido y k un entero (instancia del problema de Clique). Se construye un grafo n-partito G' para el problema de diseño de proteínas como sigue:

1. Particiones de G'
   - Cada vértice $v_i \in V$ corresponde a una partición V_i en G'.
   - Cada partición P_i tiene dos vértices:
     - a_i (representa ïncluir v_i en el clique"), con peso 0.
     - b_i (representa .[ex]cluir v_i"), con peso 1.
2. Aristas en G'
   - Para cada par de particiones V_i, V_j $(i \neq j)$:
     - Si $\langle v_i, v_j \rangle \in E$, la arista $\langle a_i, a_j \rangle$ tiene peso 0.
     - Si $\langle v_i, v_j \rangle \notin E$, la arista $\langle a_i, a_j \rangle$ tiene peso 1.
   - Todas las aristas que involucran algún b_i tienen peso 0.

3.1.2. Equivalencia de soluciones

Teorema. *G tiene un clique de tamaño k si y solo si G' tiene un subgrafo inducido de peso menor o igual que 0.*

Demostración

($\Rightarrow$)

Seleccionando los vértices a_i correspondientes a los k nodos del clique en G, el subgrafo inducido en G' cumple que:

- Los vértices a_i tienen peso 0.
- Las aristas entre ellos tienen peso 0 (porque forman un clique en G).

Por tanto, el peso total es de 0.

($\Leftarrow$)

- Todos los vértices del subgrafo deben ser a_i (ya que cualquier b_i añadiría peso).
- Las aristas entre estos a_i deben tener peso 0, lo que implica que $\langle v_i, v_j \rangle \in E$ para todo par.

Esto corresponde a un clique de tamaño k en G.