5) Inteligencia Artificial aplicada al análisis de secuencias biológicas y Clustering con ADN

La Inteligencia Artificial (IA) permite analizar secuencias de ADN identificando patrones, mutaciones y agrupando secuencias similares. Esto es útil para medicina, evolución y genómica moderna.

1. Representación de secuencias

El ADN debe convertirse a números para ser procesado por modelos de IA:

• One-hot encoding: vectores binarios para A,T,C,G.

• K-mers: frecuencias de subsecuencias como AA, AT, CG, etc.

• Embeddings: representaciones continuas aprendidas.

2. Modelos supervisados

Se utilizan para clasificar secuencias (normal/mutada, tipo de gen, especie):

• Regresión logística

• SVM

• Random Forest

3. Redes neuronales aplicadas

• CNN: detectan motivos biológicos.

• RNN/LSTM: capturan dependencias secuenciales.

• Transformers: estado del arte en análisis de secuencias largas.

4. Detección automática de mutaciones

Se usan autoencoders y modelos de anomalías que comparan secuencias reconstruidas con las reales.

5. Generación de secuencias

Modelos generativos producen ADN sintético o simulan mutaciones.

6. Clustering con K-means en ADN

K-means se usa para análisis no supervisado, agrupando secuencias según sus características:

• Familias génicas

• Variantes virales

• Perfiles tumorales

• Regiones funcionales del genoma

Las secuencias se vectorizan con k-mers, frecuencias o entropía. Luego K-means agrupa en clusters minimizando la distancia intra-cluster.

Función objetivo:

$J = \Sigma \Sigma ||x - \mu_i||^2$

Donde x es el vector de una secuencia y μ_i es el centroide del cluster.

Aplicaciones:

• Identificación de subtipos de cáncer a partir del ADN tumoral.

• Agrupamiento de bacterias en estudios de microbioma.

• Detección de regiones conservadas o repetitivas del genoma.

Conclusión:

La IA y el clustering permiten estudiar el ADN como datos textuales complejos. Su aplicación es clave en medicina de precisión, evolución y análisis masivo de datos genómicos.