Universidad Católica San Pablo

Informe del laboratorio 3

Computer Science — Biología Molecular Computacional

Harold Alejandro Villanueva Borda

---

1. Introducción

El presente informe busca analizar la consistencia de los resultados de un algoritmo de agrupamiento filogenético aplicado tanto a secuencias de ADN como a apellidos, con el fin de evaluar si los patrones obtenidos reflejan posibles parentescos entre las personas. Se emplean dos métodos de agrupamiento filogenético comúnmente utilizados en biología: UPGMA (Unweighted Pair Group Method with Arithmetic Mean) y Neighbor Joining (NJ). Los resultados serán interpretados para determinar si las distancias calculadas entre secuencias o apellidos sugieren algún tipo de parentesco entre los individuos.

2. Lógica del código

El código proporcionado se estructura en varias funciones clave:

- Cálculo de matriz de distancias: Se define una función $\text{distance}_matrix(sequences)quecalculalamatrizded$

- Implementación de UPGMA y Neighbor Joining: Ambos algoritmos crean un árbol filogenético a partir de la matriz de distancias. UPGMA es un algoritmo jerárquico basado en la suposición de que las tasas de mutación son constantes, mientras que Neighbor Joining es más flexible, permitiendo tasas de evolución diferentes.

- Visualización del árbol: La función $\text{visualize}_tree()generaunarepresentaciongráficadelárbolfilogenéticoaparti$

- Distancia entre apellidos: Para los apellidos, la función $\text{surname}_distance(surname1, surname2)calculaunadist$

- Pruebas: Se ejecutan pruebas con una lista de secuencias de ADN y otra con apellidos. Los resultados son visualizados mediante árboles filogenéticos generados con ambos métodos (UPGMA y Neighbor Joining).

3. Experimentos

Se realizaron dos experimentos: uno con secuencias de ADN y otro con apellidos.

1. Análisis de Secuencias de ADN: Las secuencias utilizadas en el análisis son:
   - ATTGCCATT
   - ATGGCCATT
   - ATCCAATTTT
   - ATCTTCTT
   - ACTGACC

   Se calculó una matriz de distancias entre las secuencias utilizando la distancia de Hamming. Posteriormente, se aplicaron los algoritmos UPGMA y Neighbor Joining para construir los árboles filogenéticos correspondientes.

2. Análisis de Apellidos: Los apellidos considerados son: - García López
   - Martínez Rodríguez
   - Fernández González
   - López Sánchez
   - González Pérez
   - Rodríguez Martín

   La distancia entre los apellidos se calculó utilizando una métrica de diferencia de caracteres, obteniendo una matriz de distancias. Luego, se construyeron árboles filogenéticos con ambos algoritmos para visualizar las relaciones entre los apellidos.

4. Resultados

1. **Árbol Filogenético de Secuencias de ADN:** Tanto el algoritmo UPGMA como el Neighbor Joining generaron árboles que mostraron relaciones esperadas entre las secuencias de ADN. Las secuencias más similares en términos de número de diferencias formaron agrupaciones cercanas en los árboles. El UPGMA asumió una tasa constante de mutación y agrupó las secuencias con esa premisa, mientras que Neighbor Joining ajustó mejor las secuencias permitiendo diferentes velocidades de cambio.

   - **Matriz de Distancias (Hamming Distance):** El análisis de las secuencias de ADN generó una matriz de distancias basada en la cantidad de posiciones en las que las secuencias difieren. A continuación se muestra una representación simplificada de la matriz de distancias:

     - Las secuencias ATTGCCATT y ATGGCCATT son las más cercanas, con solo una diferencia en sus posiciones.

ATTGCCATT	ATGGCCATT	ATCCAATTTT	ATCTTCTT	ACTGACC
0	1	5	4	5
1	0	4	3	6
5	4	0	3	6
4	3	3	0	5
5	6	6	5	0

Cuadro 1: Tabla de distancias

- Las secuencias ATCCAATTTT y ATCTTCTT también muestran una proximidad relativamente cercana, con 3 diferencias.

- **Árbol Filogenético con UPGMA:** El algoritmo UPGMA construyó un árbol filogenético que agrupa las secuencias más similares bajo la premisa de una tasa de mutación constante. El árbol resultante muestra:
  - ATTGCCATT y ATGGCCATT agrupados muy cerca, lo que refleja su baja distancia genética.
  - Las secuencias ATCCAATTTT y ATCTTCTT también forman un subgrupo dentro del árbol, reflejando su similitud.
  - La secuencia ACTGACC queda más alejada, ya que tiene una mayor distancia con respecto a las demás secuencias.
- **Árbol Filogenético con Neighbor Joining:** El método Neighbor Joining también mostró resultados similares en cuanto a la agrupación de las secuencias más cercanas. Sin embargo, este algoritmo permitió tasas de mutación variables, ajustando mejor la distancia entre las secuencias. En este árbol:
  - Se observan agrupaciones similares a las de UPGMA, pero con ramas de diferentes longitudes que representan mejor las distancias entre las secuencias.
  - ACTGACC sigue estando más distante, pero las otras secuencias mantienen una estructura jerárquica que refleja bien las relaciones de parentesco.

2. **Árbol Filogenético de Apellidos:** En el análisis de apellidos, los árboles generados por ambos métodos presentaron agrupamientos interesantes, destacando la similitud entre "García Lópezz "López Sánchez", lo que tiene sentido considerando la presencia de "López.en ambos apellidos. Sin embargo, no todos los agrupamientos reflejan necesariamente parentescos biológicos, ya que la distancia entre apellidos solo captura similitudes en caracteres.

   - **Matriz de Distancias (Caracteres Diferentes):** En el análisis de los apellidos, se calculó la distancia entre apellidos en función de las diferencias en sus caracteres. La matriz de distancias es la siguiente:

García López	Martínez Rodríguez	Fernández González	López Sánchez	González Pérez	Rodríguez Martín
0	8	7	4	7	8
8	0	9	7	6	5
7	9	0	8	5	8
4	7	8	0	9	8
7	6	5	9	0	7
8	5	8	8	7	0

Cuadro 2: Tabla de distancias

- Los apellidos García López y López Sánchez tienen la menor distancia (4), lo cual tiene sentido, ya que ambos comparten López como uno de los componentes.
- Fernández González y González Pérez también muestran cierta proximidad, con una distancia de 5, probablemente por la coincidencia en González.

- **Árbol Filogenético con UPGMA:** En el árbol resultante de UPGMA:
  - Los apellidos García López y López Sánchez se agrupan como los más cercanos.
  - Los apellidos González Pérez y Fernández González también se agrupan debido a su similitud.
  - Los demás apellidos están más separados, lo que refleja una mayor diferencia en caracteres.

- **Árbol Filogenético con Neighbor Joining:** Neighbor Joining ajustó mejor las distancias entre los apellidos:
  - Los grupos son similares a los de UPGMA, pero con mayor variabilidad en la longitud de las ramas, indicando que algunas distancias entre apellidos son más representativas.
  - Se mantiene el agrupamiento entre García López y López Sánchez, así como entre Fernández González y González Pérez.

5. Interpretación de Resultados

Los árboles filogenéticos obtenidos para las secuencias de ADN mostraron una clara consistencia con la hipótesis de parentesco, dado que las secuencias más similares fueron agrupadas cercanamente. En cambio, los resultados de los apellidos, si bien reflejan cierta similitud en nombres, no pueden ser interpretados directamente como evidencia de parentesco, ya que los apellidos, aunque pueden ser heredados, no necesariamente se relacionan con la cercanía genética.

6. Conclusión

Los resultados sugieren que los algoritmos filogenéticos son efectivos para detectar parentescos cuando se utilizan datos genéticos, como las secuencias de ADN. Sin embargo, la aplicación a apellidos es limitada, ya que las similitudes entre apellidos no siempre se corresponden con relaciones de parentesco biológico. Por tanto, el análisis de apellidos puede proporcionar algunas indicaciones, pero no puede reemplazar métodos más directos como el análisis de ADN para determinar relaciones familiares.

7. Implementación en Github

El código fuente del análisis se encuentra disponible en GitHub: GitHub.