Alineación de secuencias genéticas mediante el algoritmo de Needleman–Wunsch

Docente:

Camilo Argoty

Alumno:

Gustavo Viñas

Descripción General

En este ejercicio se aplicarán conceptos fundamentales de algoritmos y estructuras de datos en el contexto de la bioinformática, específicamente en la alineación global de secuencias de nucleótidos.

El objetivo es comprender cómo los algoritmos permiten comparar secuencias biológicas para identificar similitudes, mutaciones y relaciones evolutivas.

Parte 1: Conceptos Teóricos

El estudiante deberá investigar y redactar un documento (1–2 páginas) explicando:

1. Qué es una secuencia de nucleótidos y por qué su comparación es importante en biología molecular.
2. Qué es una alineación de secuencias, diferenciando entre:
   Alineación global y alineación local.
   Coincidencias (matches), desajustes (mismatches) y huecos (gaps).
3. El modelo de puntuación usado para alinear secuencias:
   Reglas de puntuación para coincidencias y desajustes.
   Penalización por apertura y extensión de huecos.
4. Una descripción conceptual del algoritmo de Needleman–Wunsch, incluyendo:
   a. Construcción de la matriz de programación dinámica.
   b. Ecuación de recurrencia.
   c. Proceso de traceback para recuperar el alineamiento óptimo.

El documento debe incluir al menos una referencia bibliográfica confiable.

1. Secuencia de nucleótidos y su importancia

Una secuencia de nucleótidos es una cadena lineal de unidades monoméricas llamadas nucleótidos, que son los bloques de construcción de los ácidos nucleicos (ADN o ARN).
En el ADN, los nucleótidos contienen una de las cuatro bases nitrogenadas: Adenina (A), Guanina (G), Citosina (C) y Timina (T).
En el ARN, la Timina (T) se reemplaza por Uracilo (U).
La comparación de secuencias de nucleótidos (y aminoácidos) es fundamental porque la secuencia determina la función, es decir, que el orden lineal específico de los monómeros en una molécula biológica es lo que dicta su estructura tridimensional y, por ende, el rol químico y biológico que desempeña en la célula.
Esta metodología es esencial, ya que, entre otros beneficios:

- Permite identificar similitudes entre genes o proteínas de diferentes organismos, lo que sugiere una función compartida y un ancestro común (relaciones evolutivas).
- Facilita la detección de mutaciones que pueden estar relacionadas con enfermedades.
- Ayuda a predecir la estructura y la función de nuevas secuencias.

2. Alineación de secuencias

La alineación de secuencias es una poderosa herramienta capaz de revelar los patrones y funciones de los genes. Si dos regiones genéticas son similares o idénticas, el alineamiento de secuencias puede demostrar los elementos conservados o diferencias entre ellas.
La alineación de secuencias representa el método de comparación de dos o más cadenas genéticas, como ADN o ARN. Estas comparaciones ayudan con el descubrimiento de puntos en común genéticos y con el rastreo (implícito) de la evolución de las hebras.
Hay dos tipos principales de alineación:

- Alineación global: un intento de alinear cada elemento de una cadena genética, más útil cuando las hebras genéticas consideradas son de aproximadamente el mismo tamaño. La alineación global también puede terminar en brechas.
- Alineación local: un intento de alinear regiones de secuencias que contienen motivos de secuencia similares dentro de un contexto más amplio.

Componentes del Alineamiento:

- Coincidencias (*matches*): ocurren cuando los caracteres en la misma posición de las secuencias alineadas son idénticos (e.g., A/A o G/G).
- Desajustes (*mismatches*): ocurren cuando los caracteres son diferentes (e.g., A/T o C/G).
- Huecos (*gaps*): se insertan para compensar las regiones donde un carácter está presente en una secuencia, pero no en la otra, lo que representa una posible deleción o inserción (eventos de mutación). Se representan con un guion ('-').

3. Modelo de puntuación

Para evaluar la calidad de una alineación, se asignan puntuaciones numéricas a las coincidencias, desajustes y huecos.

- Reglas de puntuación para coincidencias y desajustes:
  - Coincidencia (*match*): se asigna una puntuación positiva (ej. +1). Cuanto mayor sea la puntuación, más deseable es el *match*.
  - Desajuste (*mismatch*): se asigna una puntuación negativa (ej. -1). Esto penaliza el desacuerdo entre caracteres.

- Penalización por apertura y extensión de huecos:
  - *Gap penalty*: una puntuación negativa que se resta cada vez que se inserta un hueco.
  - En modelos más sofisticados, la penalización se divide en:
    - Penalización de apertura: un costo inicial alto por comenzar un nuevo *gap*.
    - Penalización de extensión: Un costo más bajo por cada nucleótido adicional que se extiende en el *gap* ya abierto.

4. Algoritmo de Needleman-Wunsch

El algoritmo de Needleman-Wunsch es un método de programación dinámica que garantiza encontrar el alineamiento global óptimo entre dos secuencias, maximizando la puntuación total.

Construcción de la matriz de programación dinámica

El algoritmo construye una matriz (a menudo llamada *H* o *S*) donde:

- Las filas corresponden a los caracteres de la primera secuencia (S_1).
- Las columnas corresponden a los caracteres de la segunda secuencia (S_2).
- Cada celda (i, j) de la matriz almacena la máxima puntuación posible para alinear el prefijo de S_1 de longitud i con el prefijo de S_2 de longitud j.

Inicialización:

La primera fila (i=0) y la primera columna (j=0) se inicializan con puntuaciones de hueco acumuladas.

- H(i, 0) = i x Penalización de Gap
- H(0, j) = j x Penalización de Gap

Ecuación de recurrencia

Cada celda subsiguiente H(i, j) se calcula tomando el valor máximo de tres posibilidades, representando las tres operaciones que podrían haber llevado a esa alineación:

$$H(i,j) = max \begin{cases} H(i-1,j-1) + Score(S_1[i], S_2[j]) \ (Diagonal: Match\ o\ Mismatch) \\ H(i-1,j) + Penalización\ de\ Gap \ \ (Superior: Gap\ en\ S_2) \\ H(i,j-1) + Penalización\ de\ Gap \ \ (Izquierda: Gap\ en\ S_1) \end{cases}$$

Donde:

- H(i-1, j-1) es la puntuación de la alineación de prefijos anterior más la puntuación por alinear los caracteres $S_1[i]$ y $S_2[j]$.
- H(i-1, j) es la puntuación de la alineación anterior más la penalización por insertar un hueco en la segunda secuencia (S_2).
- H(i, j-1) es la puntuación de la alineación anterior más la penalización por insertar un hueco en la primera secuencia (S_1).

Proceso de *traceback* para recuperar el alineamiento óptimo

Una vez que la matriz está completamente llena, el puntaje final del alineamiento es el valor de la celda en la esquina inferior derecha: H(M, N), donde M y N son las longitudes de las secuencias.

Para recuperar el alineamiento *real*, se realiza el proceso de traceback (rastreo):

1. Se comienza en la celda final H(M, N).
2. En cada celda (i, j), se rastrea el camino de vuelta a la celda que proporcionó el valor máximo según la ecuación de recurrencia:
   - Si el máximo vino de H(i-1, j-1) (diagonal): se alinean $S_1[i]$ con $S_2[j]$.
   - Si el máximo vino de H(i-1, j) (superior): se alinea $S_1[i]$ con un hueco (-).
   - Si el máximo vino de H(i, j-1) (izquierda): se alinea un hueco (-) con $S_2[j]$.

3. El proceso termina al llegar a la celda H(0, 0). El camino inverso reconstruido representa el alineamiento global con la máxima puntuación.

Referencias

(2016). En K. e. al., *Biología Computacional - Genomas, Redes y Evolución* (págs. 42-66). Obtenido de https://espanol.libretexts.org/Bookshelves/Biologia/Biologia_Computacional/Libro%3A_Biolog%C3%ADa_Computacional_-_Genomas%2C_Redes_y_Evoluci%C3%B3n_(Kellis_et_al.)

(2016). En P. M. Kellis, *Computational Biology: Genomes, Networks, Evolution* (págs. 25-44). Obtenido de https://ocw.mit.edu/ans7870/6/6.047/f15/MIT6_047F15_Compiled.pdf