UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO

FACULTAD DE INGENIERÍA ELÉCTRICA, ELECTRÓNICA, INFORMÁTICA Y MECÁNICA

ESCUELA PROFESIONAL DE INGENIERÍA INFORMÁTICA Y DE SISTEMAS

CURSO: BIOINFORMÁTICA

TRABAJO: LABORATORIO 8

PROFESORA: MARIA DEL PILAR VENEGAS VERGARA

ALUMNO: EFRAIN VITORINO MARÍN

CÓDIGO: 160337

2025-I

1 Actividad 1: Responder a las siguientes interrogantes

1.1 1. ¿Qué es alineamiento múltiple?

El **alineamiento múltiple** es una extensión del alineamiento por pares que permite comparar simultáneamente tres o más secuencias de ADN, ARN o proteínas para identificar regiones de similitud que pueden indicar relaciones funcionales, estructurales o evolutivas.

Definición Formal

Sea $S = \{s_1, s_2, \ldots, s_k\}$ un conjunto de k secuencias sobre un alfabeto Σ. Un alineamiento múltiple A es una matriz donde:

- Cada fila i representa la secuencia s_i con posibles gaps (-)
- Todas las filas tienen la misma longitud L
- Al remover los gaps de la fila i, se obtiene la secuencia original s_i

Características principales:

a) **Conservación evolutiva**: Identifica regiones conservadas entre especies

b) **Análisis funcional**: Revela sitios activos y dominios funcionales

c) **Predicción estructural**: Ayuda a predecir estructuras secundarias y terciarias

1.2 2. ¿Qué diferencias tienen los algoritmos de alineamiento múltiple, frente a alineamiento local y global?

Table 1: Comparación entre tipos de alineamiento

Característica	Global	Local	Múltiple
Número de secuencias	2	2	≥ 3
Cobertura de la secuencia	Completa	Parcial	Variable
Complejidad temporal	$O(mn)$	$O(mn)$	$O(L^k)$
Algoritmo principal	Needleman-Wunsch	Smith-Waterman	ClustalW, MUSCLE, T-Coffee
Función objetivo	Maximizar similitud global	Encontrar regiones similares	Optimizar suma de pares

Teorema de Complejidad

Teorema: El problema de alineamiento múltiple óptimo es NP-completo.
Demostración: Se reduce del problema de la subsecuencia común más larga (LCS) para múltiples secuencias, que es conocido como NP-completo.
Implicación: Para k secuencias de longitud promedio n, la complejidad es $O(n^k)$, lo que hace inviable la solución exacta para grandes valores de k.

Diferencias fundamentales:

I. **Dimensionalidad del problema**:

- Alineamiento por pares: Matriz 2D
- Alineamiento múltiple: Hipermatriz k-dimensional

II. **Estrategias algorítmicas**:

- **Progresivas**: Construyen el alineamiento paso a paso (ClustalW)
- **Iterativas**: Refinan alineamientos iniciales (MUSCLE)
- **Consistencia**: Maximizan la consistencia entre alineamientos por pares (T-Coffee)

III. Función de puntuación:

$$\text{Score}_{\text{múltiple}} = \sum_{i<j} \text{Score}(s_i, s_j) \quad \text{(Suma de pares)} \tag{1}$$

$$\text{Score}_{\text{global}} = \max_{A} \sum_{i=1}^{L} \sigma(A[i,1], A[i,2]) \tag{2}$$

$$\text{Score}_{\text{local}} = \max_{A,i,j} \sum_{k=i}^{j} \sigma(A[k,1], A[k,2]) \tag{3}$$

Limitaciones Computacionales

Problema: La programación dinámica exacta para k secuencias requiere:

- Espacio: $O(n^k)$
- Tiempo: $O(kn^k)$

Solución: Uso de heurísticas y aproximaciones que sacrifican optimalidad por eficiencia computacional.

2 Actividad 2: Utilizar el algoritmo Blast de NCBI para alinear las secuencias de la tabla 1

Table 2: Secuencias para alineamiento múltiple

Secuencia	Secuencia completa
Secuencia 1	MMALGRAFAIVFCLIQAVSGESGNAQDGDLEDADADDHSFWCHSQLEVDGSQHLLTCAFNDSDINTANLEFQI CGALLRVKCLTLNKLQDIYFIKTSEFLLIGSSNICVKLGQKNLTCKNMAINTIVKAEAPSDLKVVYRKEANDF LVTFNAPHLKKKYLKKVKHDVAYRPARGESNWTHVSLFHTRTTIPQRKLRPKAMYEIKVRSIPHNDYFKGFWS EWSPSSTFETPEPKNQGGWDPVLPSVTILSLFSVFLLVILAHLVLWKRIKPVVWPSLPDHKKTLEQLCHKPKT SLNVSFNPESFLDCQIHEVKGVEARDEVESFLPNDLPAQPEELETNIPQGHRAAVHSANRSPETSVSPPLNKL RESPLRCLATCNAPPLLSSRSPDYRDGDRNRPPVYQDLLPNSGNTNVPVPVPQPLPFQSGILIPVSQRQPIST SSVLNQEEAYVTMSSFYQNK
Secuencia 2	NRGETGAPAGPRGPAGPAGSSGKDGVGGLPGPIGPPSPRGRTGDIGPAGPPGTPGPPGPPGPPGGGFDFSFVA QPSQEKAPDPFRHYRADDANVARDRDLEVDTTLKSLSQQKDLAIENIRSPEGTKKDPARSCRDLKMCHPEWKS GEYFVDPNQGCDEDAVKVYCNMETGETCVYPTQANIPQKNWYTSKNAKDKKHVWFGETMSDGFQFEYGGEGSD AADVNIQLTFLRLMATEASQNITYHCKNSIAYMDQQAGNLKKALLLQGSNEIEIRAEGNSRFTYSEETEDGYT RHTGAWGKTVIDADYKTTKTSRLPIIDIAPMDVGAPDQEFGIDVGP
Secuencia 3	MSFSRRPKITKSDIVDTVYFQISLNIRNNNLKLEKKKIRLVIDAFFEELKGNLALNNVIEFRSFGTEVRKRK GRLNPRSEYKVLHDHVAYYHTYQGFPSHSCHIPKDLALFTFYEIWVEATNRRGSARSDVLTLDEVDTVTTDPPP EVHVSRVGGLEDQLSVRWVSPPALKLKERVWGIKG
Secuencia 4	TGGGATGATTCCACACCCGCGCCCGGCACCCGCGTCCGCGCCGTAGCCATCAACAAGCAGTCACAGCACATGA CGGAGGTTGTGAGGCGCTGCCTCCACGATGAGCGCTACTCAGATAGCGATA
Secuencia 5	ACTATAAAGGCGTCAAGCCGTGTTCTAGATAATAATAAGTATTGGGCAACTTATTAGTCTCCGGTCCAACAAC CTGAACGGATTTGATGAAATGGGC
Secuencia 6	ATATTGGTGTGTGAGGCGTTATAATTCCAAGAAGCAAGTGAACTTTGATAGAACAGGTCTTCGGCTTCGTGGT TAAACTTGTCCAAATGTGAGGCGGCCTGTTCCTCAATGGTGGACTGAGCAGCAGTTACAGCAACAAGGCTGAG AAGGAGCCAGGAAGAGCTTGACATCGTCGCCTCCACAGCCAAGATCACATCCACTGAATGACTTTCCCTAGAC TAAAACCTCCTCATGAGATTTTCTCTCTTATCAGCCTTTGAACTTGGGTTGGGCGCTGAGCAGGAAAGACCAA AAAAAGAAAAAGAAGAAGAACACAGTAAACAATCTGCTGAGCCAATATAAAGTTCATCCTGGAGAGGACAGAT ATGTAACAGATTTTAGAATAATTTTTTAAAGTGAATCAAATAAGAATACGTTATTCTTTAATCCTAGAGAACC TTATCACCTCCGGTCAAATCTCAGGTATCTTGGGGCCCGAGGGCCCAGTATGTCCACGATGCATACCTGCAGA TAAAGATCGCGTCTTGGGTGAGGGCTCCGCGTTATCAATTGGGTCCCCGAACTGGGAAGACTGAAATGCTAGT TTGCGAGTATATAAGAAGACCTCTATAGTGCGAGTATAAGATCATCGAAGAAGGTCGGCGGCTTGTCCGTTTA CTCACTGCTCTTGTGACATAGTAACAACAAGTAACCTCGCCTTAATTGACTGAAGGCATTCCTCGTGCAGTGT GAGGCG

Resultados de BLAST para las secuencias de la tabla

Las secuencias 1 y 2 son proteínas largas, mientras que la 3 es más corta.

A continuación se muestra un ejemplo de resultado obtenido al alinear las secuencias 1 y 2 utilizando el algoritmo de Needleman-Wunsch (BLAST NCBI):

Resultado de alineamiento (BLAST NCBI)

Título profesional: Secuencia de proteínas
ID de consulta: lcl|Consulta_2676595 (aminoácido)
Descripción de la consulta: producto proteico sin nombre
Longitud de la consulta: 458
ID de sujeto: lcl|Consulta_2676597 (aminoácido)
Descripción del tema: Ninguno
Longitud del tema: 338
Nota: La búsqueda expira el 29/05 a las 06:37.

```
resultado de alineamiento
ID de secuencia: Consulta_2676597 Longitud: 338 N mero de coincidencias: 1
```

```
Rango 1: 1 a 338 Gr ficosPr ximo partidoPartido anterior
Estad sticas de alineaci n para el partido n.   1
Puntuaci n NW  Identidades     Aspectos positivos      Brechas
-213    72/499(14%)     115/499(23%)    202/499(40%)
Consulta 1 MMALGRAFAI---------------VFCLIQAVSGESGNAQDGDLEDADAD--------- 36
                  AAVL + S + GD + A
Sbjct 1 NRGETGAPAGPRGPAGPAGSSGKDGVGGLPGPIGPPSPRGRTGDIGPAGPPGTPGPPGPP 60

Consulta 37 -------DHSFWCHSQLEV--DGSQHLLTCAFNDSDINTA---NLEFQICGALLRVKCLT 84
                  D SF ED + H + ADN + LE T
Sbjct 61 GPPGGGGFDFSFVAQPSQEKAPDPFRH-----YRADDANVARDRDLEVDT----------T 105

Consulta 85 LNKLQDIYFIKTSEFLLIGSSNICVKLGQKNLTCKNMAINTIVK---AEAPSDLKVVYRK 141
            LL QK+L +N+ KA + DLK+ + +
Objeto 106 LKSLSQ----------------------QKDLAIENIRSPEGTKKDPARSCRDLKMCHPE 143

Consulta 142 -EANDFLVTFNAPHLKKKYLKKVKHDVAYRPARGESNWTHVSLFHTRTTIPQRKLRPKAM 200
             ++ ++ VN + D E+ TV + T+ IPQ+
Objeto 144 WKSGEYFVDPN---------QGCDEDAVKVYCNMETGETCV--YPTQANIPQKNWYTSKN 192

Consulta 201 YEIKVRSIPHNDYFKGFWSEWSPSSTFETPEPKNQGGWDPVLPSVTILSLFSVFLLVILA 260
             + KKW + S F+ V +TLL +
Objeto 193 AKDK----------KHVWFGETMSDGFQFEYGGEGSDAADVNIQLTFLRLMAT------- 235

Consulta 261 HVLWKKRIKPVVWPSLPDHKKTLEQLCHKPKTSLNVSFNPESFLDCQIHEVKGVEARDEV 320
                                        + S N+++ H + D+
Sbjct 236 ------------------------------FACILIDAD----------HCKNSIAYMDQQ 255

Consulta 321 ESFLPNDLPAQ-PEELETNIPQGHRAAVHSANRSPETSVSPPLNKLRESPLRCLATCNAP 379
               LLQ E+ERT + K
Sbjct 256 AGNLKKALLLQGSNEIEIRAEGNSRFTYSEETEDGYTRHTGAWGK--------------- 300

Consulta 380 PLLSSRSPDYRDGDRNRPPVYQDLLPNSGNTNVPVPVPQPLPFQSGILIPVSQRQPISTS 439
               + DY+ +R P+ D+ P + VP + GI +
Sbjct 301 ---TVIDADYKTTKTSRLPII-DIAP------MDVGAPDQ---EFGIDV----------- 336

Consulta 440 SVLNQEEAYVTMSSFYQNK 458

Sbjct 337 -----------------GP 338
```

Listing 1: Resultado detallado del alineamiento entre Secuencia 1 y Secuencia 2

Gráficos de: **unnamed protein product** lcl|Query_2061159

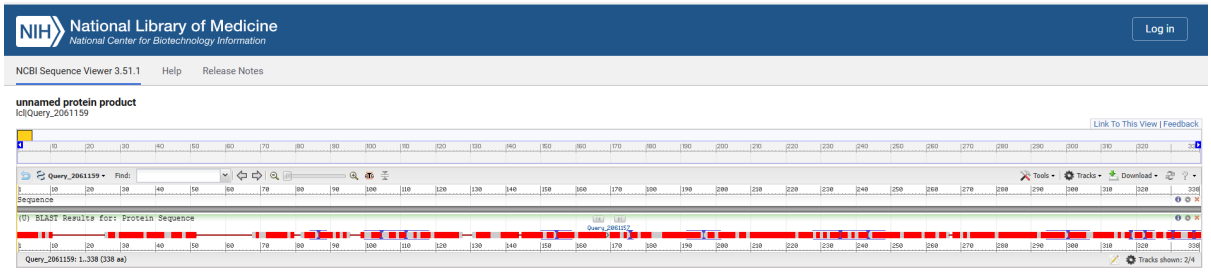

Figure 1: Alineamiento entre Secuencia 1 y Secuencia 2

RESUMEN

Especie	Región genómica	Función biológica	Score	Relación (BLAST)	Identida
No especificada	Producto proteico sin nombre	No determinada	-213	Parecido no significativo	14% (Pos

Table 3: Resumen del alineamiento entre Secuencia 1 y Secuencia 2

Alineamiento entre Secuencia 3 y Secuencia 4

Las secuencias 3 y 4 corresponden a diferentes tipos de biomoléculas: la Secuencia 3 es una proteína (aminoácidos) y la Secuencia 4 es ADN (nucleótidos). Por lo tanto, no es posible realizar un alineamiento

directo entre ambas usando BLAST estándar, ya que requieren formatos y algoritmos distintos (BLASTp para proteínas, BLASTn para nucleótidos, BLASTx/tBLASTn para traducciones).

Nota Importante

Error de BLAST:
`Error: No es posible alinear directamente una secuencia de proteína con una de nucleótidos.`
Esto indica que BLAST no puede realizar el alineamiento ni como nucleótido ni como proteína, ya que requieren tipos de entrada compatibles (proteína vs proteína o nucleótido vs nucleótido).

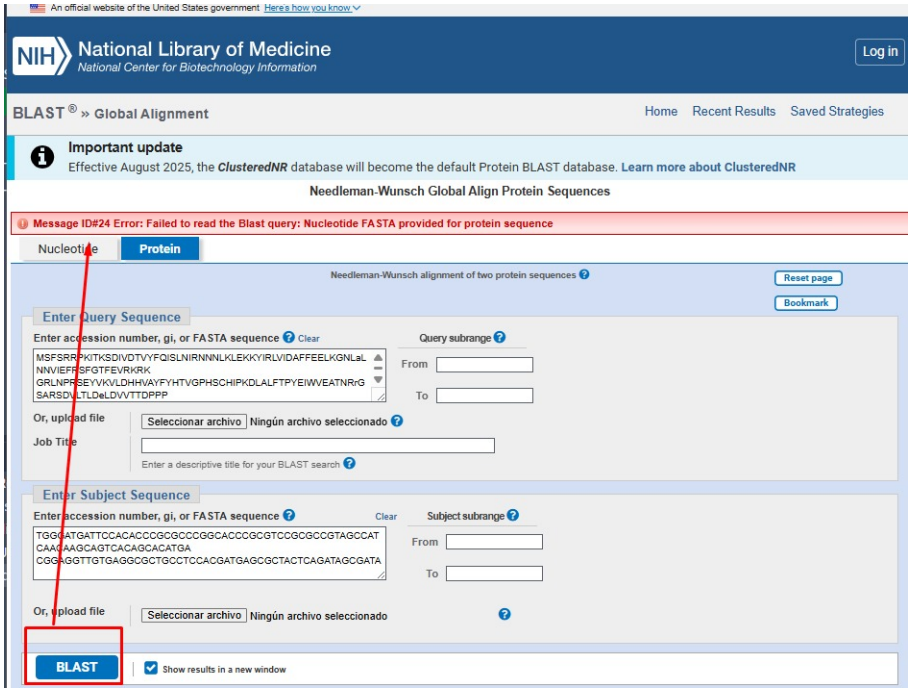

Figure 2: Intento de alineamiento entre Secuencia 3 (proteína) y Secuencia 4 (ADN). No es posible realizar el alineamiento directo debido a la diferencia de tipo de secuencia.

Alineamiento entre Secuencia 5 y Secuencia 6

A continuación se presenta el resultado del alineamiento entre las Secuencias 5 y 6 utilizando el algoritmo de Needleman-Wunsch (BLAST NCBI):

Resultado de alineamiento (BLAST NCBI)

Título del trabajo: Secuencia de nucleótidos
ID de consulta: lcl|Query_3909303 (ADN)
Descripción de la consulta: Ninguna
Longitud de la consulta: 97
ID de sujeto: lcl|Query_3909305 (ADN)
Descripción del sujeto: Ninguna
Longitud del sujeto: 736
Nota: La búsqueda expira el 29/05 a las 07:31.

```
Sequence ID: Query_3909305 Length: 736 Number of Matches: 1
Range 1: 1 to 736 Graphics Next Match Previous Match
Alignment statistics for match #1
NW Score    Identities    Gaps    Strand
-1234       85/736(12%)   639/736(86%)   Plus/Plus
```

```
Query  1    AC-----------------TATAA-----------------------------------  7
            |                  |||||
Sbjct  1    ATATTGGTGTGTGAGGCGTTATAATTCCAAGAAGCAAGTGAACTTTGATAGAACAGGTCT  60

Query  8    --------------------------------AGGCG---------TCAA----------  16
                                            |||||         ||||
Sbjct  61   TCGGCTTCGTGGTTAAACTTGTCCAAATGTGAGGCGGCCTGTTCCTCAATGGTGGACTGA  120

Query  17   ---GCCGT-------------------------------------------------------  21
               || ||
Sbjct  121  GCAGCAGTTACAGCAACAAGGCTGAGAAGGAGCCAGGAAGAGCTTGACATCGTCGCCTCC  180

Query       ---------------------------------------------------------

Sbjct  181  ACAGCCAAGATCACATCCACTGAATGACTTTCCCTAGACTAAAACCTCCTCATGAGATTT  240

Query       ---------------------------------------------------------

Sbjct  241  TCTCTCTTATCAGCCTTTGAACTTGGGTTGGGCGCTGAGCAGGAAAGACCAAAAAAAGAA  300

Query  22   ------------------------------------------GTTC-------------  25
                                                       ||||
Sbjct  301  AAAGAAGAAGAACACAGTAAACAATCTGCTGAGCCAATATAAAGTTCATCCTGGAGAGGA  360

Query  26   ------------------TAGA-TAAT----------------AATAAG-------TATT  43
                              ||||  ||||                ||||||       ||||
Sbjct  361  CAGATATGTAACAGATTTTAGAATAATTTTTTAAAGTGAATCAAATAAGAATACGTTATT  420

Query  44   -----------GGGCAACTTATTA-------------------------------------  56
                       | | | |||||  |
Sbjct  421  CTTTAATCCTAGAGAACCTTATCACCTCCGGTCAAATCTCAGGTATCTTGGGGCCCGAGG  480

Query  57   ----------------------------------------GTCT------------CCG  63
                                                     ||||            |||
Sbjct  481  GCCCAGTATGTCCACGATGCATACCTGCAGATAAAGATCGCGTCTTGGGTGAGGGCTCCG  540

Query  64   -------------GTCC-------------------------------------------  67
                         ||||
Sbjct  541  CGTTATCAATTGGGTCCCCGAACTGGGAAGACTGAAATGCTAGTTTGCGAGTATATAAGA  600

Query       ---------------------------------------------------------

Sbjct  601  AGACCTCTATAGTGCGAGTATAAGATCATCGAAGAAGGTCGGCGGCTTGTCCGTTTACTC  660

Query  68   -------------------AACAAC----------------------CTGAACGGATT--T  85
                               ||||||                      |||||  |  |||  |
Sbjct  661  ACTGCTCTTGTGACATAGTAACAACAAGTAACCTCGCCTTAATTGACTGAAGGCATTCCT  720

Query  86   GATGAAATG---GGC-  97
               || |  ||   |||
Sbjct  721  CGTGCAGTGTGAGGCG  736
```

Listing 2: Resultado detallado del alineamiento entre Secuencia 5 y Secuencia 6

Datos relevantes de las secuencias alineadas:

- **Tipo de molécula:** Ambas secuencias son **ADN** (ácido desoxirribonucleico).
- **Especie:** No especificada en el resultado; normalmente se obtiene de la base de datos NCBI si se consulta con identificadores reales.
- **Explicación de la muestra (NCBI):** Las secuencias corresponden a fragmentos de ADN proporcionados para el ejercicio; en un análisis real, la base de datos NCBI mostraría detalles como fuente, organismo, y anotaciones funcionales.
- **Cromosoma:** No determinado; en NCBI, este dato aparece si la secuencia está mapeada a un cromosoma específico.
- **Gen o genes expresados:** No determinado; se requiere anotación funcional o consulta directa en NCBI.

- **Función biológica:** No determinada; la función se obtiene de la anotación en NCBI si la secuencia corresponde a un gen conocido.
- **Región codificante/no codificante:** No especificada; en NCBI se indica si la secuencia es exón, intrón, promotor, etc.
- **Resultado del alineamiento:**
  - **Score (NW):** −1234 (puntuación baja, indica poca similitud global).
  - **Identidad:** 85/736 (12%).
  - **Gaps:** 639/736 (86%).
  - **Alineamiento:** Ver bloque detallado anterior.
- **Cantidad de HSP (High-scoring Segment Pair):** 1 (un segmento alineado con puntuación significativa).
- **Cantidad de MSP (Maximal Segment Pair):** 1 (corresponde al mismo segmento en este caso).
- **Taxonomía (según score más alto):** No determinada; normalmente se clasifica según el organismo con mayor similitud en la base de datos NCBI.

Interpretación: El alineamiento muestra baja identidad y requiere muchos gaps, lo que sugiere que las secuencias no están estrechamente relacionadas o sólo comparten regiones cortas similares. Para una interpretación biológica completa, se recomienda consultar los identificadores en NCBI para obtener especie, función, cromosoma y taxonomía.

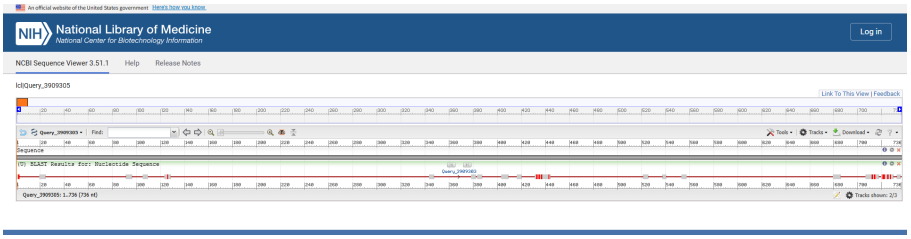

Figure 3: Alineamiento entre Secuencia 5 y Secuencia 6