Exámen 01: Búsquedas usando BLAST:
Bioinformática
Pontificia Universidad Javeriana - Cali

Prof. Luis Garreta

10 de septiembre de 2018

1. Primera Parte

- Para algunas tareas, utilizará la herramienta BLAST en la página web de NCBI http://www.ncbi.nlm.nih.gov/. Puede ubicar fácilmente la herramienta en la sección "Recursos populares". Para otras herramientas, busque en Internet la más adecuada o experimente una.
- Escriba un informe sobre sus resultados (llámelo "general_questions.pdf") y agregue cifras descriptivas cuando sea necesario. los la tarea de esta conferencia es casi idéntica a la práctica, así que te recomiendo que ya escribas sus respuestas abajo

1.1. Pregunta

1. Tienes una secuencia ATGGTGA . Crea secuencias homólogas de esa secuencia donde durante la evolución cuando se ha producido una deleción, se ha producido una inserción y se ha producido una sustitución. (por ejemplo, TGGGTGG $\rightarrow$ TGGGTG (deleción)).
2. ¿Cuál es la diferencia entre los algoritmos BLAST blastn y blastp ? ¿Qué son? ¿En que casos se usan?
3. Usted tiene una secuencia de proteína putativa EYLASLGRKHRAVGVKLSSFSTVGESLLYMLEKC que usted quiere explorar más de cerca. Use BLAST solo contra secuencias humanas. a
   a) ¿Cuál es la ID de acceso y la puntuación máxima de la secuencia de BLAST de mayor puntuación?
   b) ¿Cuál es el nombre del gen que codifica esta secuencia de proteína?
   c) ¿Se ha detectado algún supuesto dominio de conservación? Si es así, ¿cuáles son estos supuestos dominios conservados detectados?
4. Consulte la siguiente secuencia de nucleótidos usando la herramienta BLAST en la web de NCBI. BLAST contra secuencias humanas solamente y contra de la base de datos "Secuencias de ARN de referencia (refseq_rna)"

   GCTCTGGTGACCAGGACAAGGGAGGGAAGGAAGGACCCTGTGCCTGGCAAAAGTCCAGGTCGCTTCTCAGGATTTGT
   GGCACCTTCTGACTGTCAAACTGTTCTTGTCAATCTCACAGGCTCCTGGTTGTCTACCCATGGACCCAGAGGTTCTT
   TGACAGCTTTGGCAACCTGTCCTCTGCCTCTGCCATCATGGGCAACCCCAAAGTCAAGGCACATGGCAAGAAGGTGC
   TGACTTCCTTGGGAGATGCCACAAAGCACCTGGATGATCTCAAGGGCACCTTTGCCCAGCTGAGTGAACTGCACTGT
   GACAAGCTGCATGTGGATCCTGAGAACTTCAAGGTGAGTCCAGGAGATGTTTCAGCCCTGTTGCCTTTAGTCTCGAG
   GCAACTTAGACAACGGAGTATTGATCTGAGCACAGCAGGGTGTGAGCTGTTTG

   a) ¿Cuántos resultados obtienes?
   b) ¿A qué gen pertenece esta secuencia según la identidad del 100 %?
   c) Cuáles son las coordenadas del *match* en la secuencia del sujeto del mejor *match*?

d) Cuál es la identidad del segundo mejor puntaje de puntuación? Explica lo que significa y ¿Qué nucleótidos no coinciden entre sí?

5. Recuperar la secuencia de proteína de NCBI codificada por la subunidad de hemoglobina humana gamma-1 (ID de acceso es NP_000550.2) y secuencia de proteínas codificada por hemoglobina gamma A de rata (el ID de acceso es NP_742090.1). Muestra estas secuencias en tu informe. Vaya al sitio de EBI (http://www.ebi.ac.uk) y ubique las herramientas de alineación de pares de secuencia (tools PSA https://www.ebi.ac.uk/Tools/psa/)

   a) Utilice el algoritmo Needleman-Wunsch (EMBOSS Needle) para realizar una alineación global en las dos secuencias de proteínas anteriores. Copie y pegue el archivo de alineación completo (excepto los parámetros del programa) a su informe. En función de los resultados,
      - ¿Cuál es la identidad y la similitud de esta alineación con los parámetros predeterminados?
      - Hay algunos *gaps*?
      - Asigne la apertura de gap y la extensión de gap a valores muy bajas y/o muy altos y ejecutar el algoritmo de nuevo. ¿Cuál es la diferencia entre la alineación anterior y la nueva alineación. Copie y pegue los archivos de alineación completos (excepto los parámetros del programa) a su informe.

   b) Ahora use el algoritmo Smith-Waterman (EMBOSS Water) para la alineación local en el dos secuencias de proteínas arriba. Copie y pegue el archivo de alineación completo (excepto parámetros del programa) a su informe. En función de los resultados,
      - Cuál es la identidad y similitud de esta alineación con los parámetros predeterminados?
      - Hay algunos gaps?
      - Selecciona el apertura de gap y extensión de gap a penalizaciones muy bajas y / o muy altas y ejecuta el algoritmo de nuevo. ¿Cuál es la diferencia entre la alineación anterior y la nueva alineación. Copie y pegue los archivos de alineación completos (excepto los parámetros del programa) a su informe.

   c) Dibuja un *dot plot* usando las dos secuencias de proteínas en el ejercicio 5. Busca alguna herramienta de dotplot en la web.
      - Use el tamaño de ventana 1 y la identidad mínima 1 para dibujar el gráfico. Copie y pegue el plot a tu reporte
      - Cambia el tamaño de la ventana y la identidad mínima y dibuja el plot. Reporta el plot en tu informe Explica qué sucede cuando cambias los requerimientos de tamaño de la ventana y la identidad mínima.

6. Usa las secuencias LYMLEKCLGPAFTPATRAAWSQLYGAV y LYMLFTPAFTPATRAAASQL y crea un matriz de puntuación con el algoritmo Needleman-Wunsch a mano (en Excel, por ejemplo). Utilizar la Matriz BLOSUM62 para calcular los puntajes de sustitución de aminoácidos en la matriz de puntuación y penalización de gap -2. Reporte la matriz de puntuación y la alineación final.

2. Segunda Parte

El GenBank contiene más de 200 millones de secuencias, con más de 260 mil millones de nucleótidos. BLAST puede usarse para comparar una secuencia desconocida con todas las secuencias en GenBank y encontrar secuencias que coincidan. Esto puede ser útil para determinar la posible identidad de una secuencia desconocida.

En esta actividad usarás BLAST para identificar secuencias desconocidas, para esto:

1. Abre en una ventana la guía de los pasos generales para búsquedas en BLAST vistos en clase.

2. Busca el sitio de BLAST en NCBI y ábrelo en una segunda ventana.

3. Toma tres secuencias (abajo) de acuerdo a la siguiente tabla y responda las preguntas (al final)

Estudiante	Seq1	Seq2	Seq3
1	1	2	5
2	2	4	6
3	3	5	7
4	4	6	1
5	5	7	2
6	6	1	3
7	7	2	4
8	1	3	8

2.1. Secuencias

- Secuencia 1

  TCGAAATAACGCGTGTTCTCAACGCGGTCGCGCAGATGCCTTTGCTCATCAGATGCGACCGCAAC
  CACGTCCGCCGCCTTGTTCGCCGTCCCCGTGCCTCAACCACCACCACGGTGTCGTCTTCCCCGAA
  CGCGTCCCGGTCAGCCAGCCTCCACGCGCCGCGCGCGCGGAGTGCCCATTCGGGCCGCAGCTGCG
  ACGGTGCCGCTCAGATTCTGTGTGGCAGGCGCGTGTTGGAGTCTAAA

- Secuencia 2:

  GTTTATTAGTGATCATGGCTAAGTTTGCGTCCATCATCGCACTTCTTTTTGCTGCTCTTGTTCTTT
  TTGCTGCTTTCGAAGCACCAACAATGGTGGAAGCACAGAAGTTGTGCGAAAGGCCAAGTGGGACAT
  GGTCAGGAGTCTGTGGAAACAATAACGCATGCAAGAATCAGTGCATTAACCTTGAGAAAGCACGAC
  ATGGATCTTGCAACTATGTCTTCCCAGCTCACAAGTGTATCTGCTACTTTCCTTGTTAATTTATCG
  CAAACTCTTTGGTGAATAGTTTTTATGTAATTTACACAAAATAAGTCAGTGTCACTATCCATGAGT
  GATTTTAAGACATGTACCAGATATGTTATGTTGGTTCGGTTATACAAATAAAGTTTTATTCACCA

- Secuencia 3:

  CTCGAGACTAGTTCTCTCTCTCTCTCTCTCGTGCCGCATCTCACACCTGTGGATGGACGGCAGCTG
  AACCGCGGGAAACTTTCGTTCTCACTCTACCTAGATGAACTTTAGTTTATATTAAACACGCGTCGA
  CTCCCACACAAACCGTGCTCGTTTTACATCTTTGTCTCCGCTTTTGAAAACGAGAAGTTGAATTCG
  CAAGACGCAACTTTCCAGCCCCTCACTGAGCGGGCAGAGTCCGTGAAGCGATGGAGCCGTCCGTCA
  TTCCCGGTGCTGACATACCCGACCTTTACTCCATTAACCCGTTTAATGTCACTTTTCCCGACGACG
  TTTTGAGTTTCGTTCCTGATGGGAGGAACTACACCGAACCTAACCCGGTAAAGAGCCGCGGAATCA
  TCATCGCCATTTCCATCACCGCTC

- Secuencia 4:

  GACATTACGGCGACCCAGTCTCCCCCGGTGTTGTCAGTGGGACTGGGCCAGACCGCAACCATCACTT
  GTACGGCCAGTCAAAGCATCTACAGTAACCTTGCTTGGTACCAGCAGAGAGAAGGACAGAAGCCCTC
  TCTCCTGATCTATGCTGCGACAACGCGATACGAAGGAGTCTCCGAGCGATTCAGCGGCAGTGGATCA
  GGGACCAGTTTCACCCTGACAATCAGCAACGTTCAGAATGAGGATGTCGCTGACTATTACTGTCAGA
  TCGCATATTCGATCTACTCCGGTTCCGTTGTTTTCGGTGAAGGAACCAAGCTCAGACTGAGCCGT

- Secuencia 5:

  GAATTCGCGGCCGCATGGGGGAGAAGCTGCCGGTTGTGTATAAACGCTTCATCTGCTCGTTCCCGGA
  TTGTAATGCCACGTATAACAAGAACCGGAAGCTGCAGGCCCATCTGTGCAAGCACACGGGGGAGAGA
  CCGTTTCCTTGCACATATGAAGGCTGTGAGAAAGGCTTTGTGACGCTGCATCACCTGAATCGTCATG
  TGCTCTCCCACACCGGGGAGAAACCCTGCAAATGCGAAACGGAAAATTGCAATTTGGCGTTCACCAC
  AGCATCCAACATGAGGTTGCACTTCAAAAGGGCTCATTCTTCTCCGGCGCAGGTCTACGTGTGTTAT
  TTCGCAGACTGTGGCCAGCAGTTCAGGAAACATAACCAGCTAAAAATTCACCAGTATATCCATACAA
  ACCAGCAACCCTTCAAAT

- Secuencia 6:

GCCCAGCGTCTCTCGGAGGAAGCTAATTCTCAGGTTATCGCAGAGGAATCTCTTGTAGCTCGTGCTGA
GGCTACCGTTGTCCAAGCCGCCGCTCCAACCAAATCCCTTGATCTGACAACATGGAAGTATGCTGATC
TCAGAGACACTATCAACACCTCAATCGATATTGCGCTCCTGTCAGCCTGCAAGGAGGAGTTCCATCGT
CGTCTCAAGGTCTACCACGCCTGGAAGATGAAGAATAAGAAGGTTGCCGCCGGCGACAAGGGCGGACC
AGAGAGGGCTCCACAATCCATCTTTGAAAGTGCCCAACAATACAACCAGCTGGCACCCCCTCCGAAAG
CCACCAAGGCTGCCCCAGCCAATCAGAACATCCAACGCTTCTTCAGGGTGCCTTTCTCCGTGACTGGG
TCCACCGCTCAGGGTCAGATGCCCGAGAGGGGTTGGTGGTACGCCCACTTTGACGGTCAGTGGATCGC
CCGCCAGATGGAGGTACACCCCACCAAGGTCCCCGTTCTTCTGGTTGCAGGTAAAGATGATGAGAACA
TGTGTGAGATGAGTTTGGAGGAGACTGGGTTGACACGACGTCCCAACGCCGAGATCGTCGAGCGGGAG
TTTGAGGAGCCCTGGAAGCGTAGCGGCGGTCAGCAGTACCACATGGCTGCAGTACGCAACAAGCAGGC
TAGACCAACGTGGGCCACGCAGAGCTTGAA

- Secuencia 7:

AACAATTCATTTTTCCTGCTTTCCTAGAAAATTCTATAAAAGCTTCAAAATGAATTACTTGGTGATGA
TTAGTTTGGCACTTCTCTTCGTGACAGGTGTAGAGAGTGTAAAAGACGGTTATATTGTCGACGATGTA
AACTGCACATACTTTTGTGGTAGAAATGCATACTGCAACGAGGAATGTACCAAGTTGAAAGGTGAGAG
TGGTTATTGCCAATGGGCAAGTCCATATGGAAACGCCTGTTATTGCTATAAATTGCCCGATCATGTAC
GTACTAAAGGACCAGGAAGATGCCATGGCCGATAAATTATAAGATGGAATGTATCCTAAGTATCAATG
TTAAATAAATATAATCAAAAAATT

- Secuencia 8:

ACAGCAAGCGAACCGGAATTGCCAGCTGGGGCGCCCTCTGGTAAGGTTGGGAAGCCCTGCAAAGTAAAC
TGGATGGCTTTCTTGCCGCCAAGGATCTGATGGCGCAGGGGATCAAGATCTGATCAAGAGACAGGATGA
GGATCGTTTCGCATGATTGAACAAGATGGATTGCACGCAGGTTCTCCGGCCGCTTGGGTGGAGAGGCTA
TTCGGCTATGACTGGGCACAACAGACAATCGGCTGCTCTGATGCCGCCGTGTTCCGGCTGTCAGCGCAG
GGGCGCCCGGTTCTTTTTGTCAAGACCGACCTGTCCGGTGCCCTGAATGAACTGCAGGACGAGGCAGCG
CGGCTATCGTGGCTGGCCACGACGGGCGTTCCTTGCGCAGCTGTGCTCGACGTTGTCACTGAAGCGGGA
AGGGACTGGCTGCTATTGGGCGAAGTGCCGGGGCAGGATCTCCTGTCATCTCACCTTGCTCCTGCC

2.2. Preguntas

1. ¿ Cuál es el tamaño de la secuencia que se utilizó para buscar en la base de datos? Incluye las unidades.
   Fuente de Información:

   a) ¿Qué secuencia coincide con tu consulta? ¿Qué datos respaldan esta conclusión?
   Fuente de Información:

   b) ¿Qué organismo es la fuente más probable de la secuencia?
   Fuente de información:

   c) ¿Cuál es el nombre común para este organismo?
   Fuente de información:

   d) ¿Qué *phylum* contiene a este organismo?
   Fuente de información:

   e) ¿Cuál es el número de acceso para la secuencia de mejor coincidencia?
   Fuente de información:

   f) Estime el número de secuencias con un valor E menor que 0.01.
   Fuente de información:

   g) Si es posible, indique los nombres de al menos tres organismos diferentes con valores E significativos. Registre el nombre del organismo, el nombre común y el valor E.
   Fuente de información:

   h) Toma la primera secuencia coincidente en la tabla. Para esa secuencia sujeto (*subject)*, determine la longitud de la alineación, en nucleótidos, y la fracción de nucleótidos que coinciden con su secuencia de consulta.

i) Observe la alineación con la primera secuencia coincidente y determine la longitud de la alineación y la fracción de nucleótidos que coinciden con su secuencia. Haga un dibujo para representar la alineación entre las dos secuencias e incluya las posiciones de inicio y final del mapa para ambas secuencias.

```
Query   ----------------------------------
Subject ----------------------------------
```

j) Use los registros de GenBank, PubMed, Gene y UniGene para encontrar la posible función de la proteína especificada por su secuencia de ADN. Describa qué se conoce sobre el papel de esta proteína en el organismo que proporcionó el ADN.

Fuente de información:

k) ¿Se expresa esta secuencia? ¿Cómo lo sabes?

Fuente de información:

l) Si su secuencia se expresa, ¿dónde se expresa?

Fuente de información:

m) ¿Hay un tiempo específico durante el desarrollo cuando se expresa este gen?

Fuente de información:

n) ¿Se conoce algo sobre los factores que hacen que se exprese tu secuencia?

Fuente de información: