TP 5

Alumnos: Cuoco Carlos, Markon Mariano, Verdecanna Mariano

👉PARA PENSAR:¿Qué tipo de información se puede extraer de la comparación de secuencias?¿Cómo esperás que se vea en una comparación?

Al comparar dos secuencias podemos ver que cadena de aminoácidos tienen en común , lo cual nos permite encontrar cosas en comunes entre las mismas. Entendemos que podríamos extraer que parte de la secuencia intervienen en ciertas acciones del mundo real.Ejemplo si analizamos distintas secuencias pertenecientes a seres que respiran bajo el agua, podríamos encontrar que patrón de secuencia del material genético se encarga de esta tarea.
Me imagino que se resaltará las partes de las secuencias iguales, las iteraciones que sucede esto y el porcentaje de diferencia de las secuencias.

👉PARA PENSAR:¿Por qué crees que es mejor evaluar las relaciones evolutivas lejanas comparando proteínas?

Debido a que las proteínas están entre las estructuras menos compleja de un organismo tienen una cantidad finita , y se puede encontrar la información en cualquier parte del organismo(una gota de sangre por ejemplo) es mejor que utilizar otros métodos (por ejemplo ir observando las distintas adaptaciones de una especie en el entorno)

👇 RETO I: Intentemos, entonces alinear estas dos palabras, para comprender mejor el problema. Alineá en la tabla interactiva las palabras "BANANA" y "MANZANA".

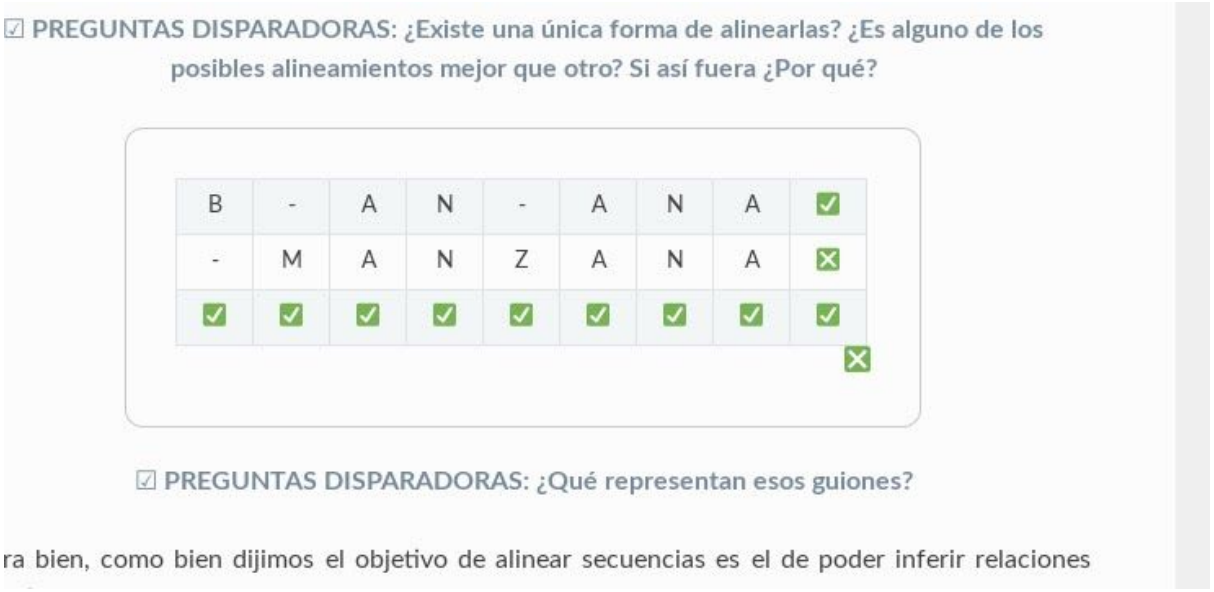
¡Tomá nota de tus observaciones y de las conclusiones que se desprendan de estas observaciones!

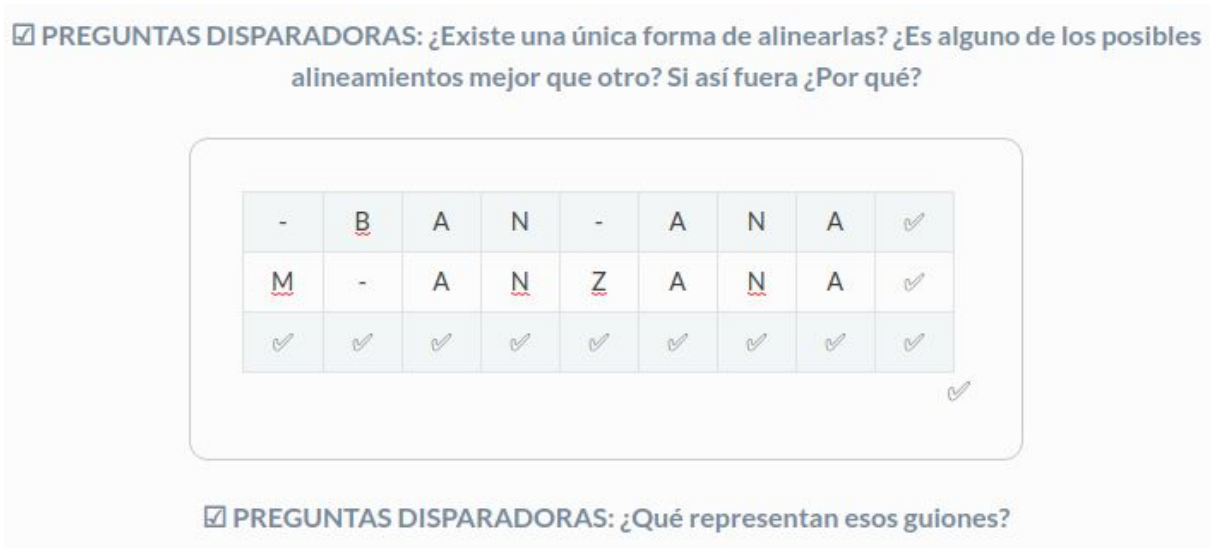
☑ PREGUNTAS DISPARADORAS: ¿Existe una única forma de alinearlas?

¿Es alguno de los posibles alineamientos mejor que otro? Si así fuera ¿Por

qué?

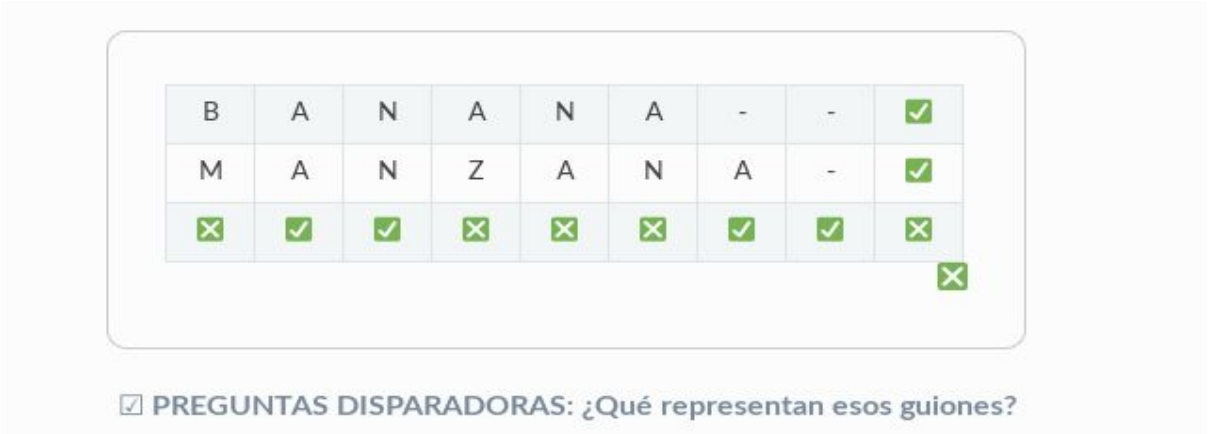
Hicimos varios pruebas y las dos opciones que más nos convencieron fueron las siguientes :





Nos dimos cuenta que existen múltiples formas de alinear 2 secuencias, no todas son igual de buenas. Cuando utilizamos un corrimiento (gap) podemos obtener un mejor alineamiento, pero si no lo ubicamos en la posición adecuada, podríamos no estar generando ninguna mejora o incluso empeorarlo.
Claramente hay mejores alineamientos, pensamos que la mayor cantidad de letras teniendo la misma posición en ambas cadenas , nos daría un alineamiento mejor.

Por ej en una foto que adjuntamos podemos ver que hay 5 casos que la letra y la posición es igual, pero en este otro alineamiento:



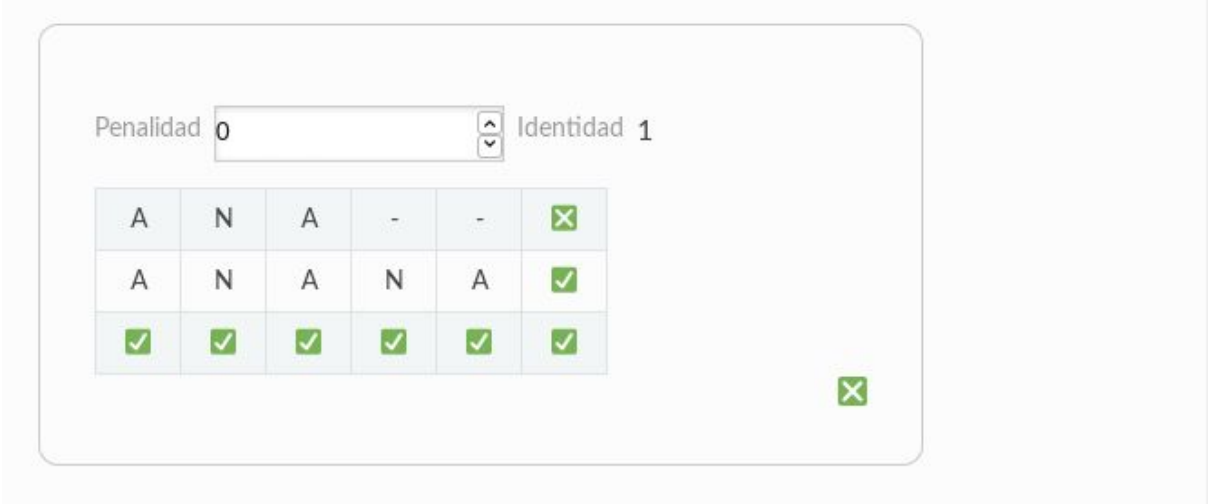
Vemos que solamentes hay 3 casos que la letra y la posición es igual en las dos cadenas.

👇 RETO II: En la siguiente tabla interactiva distintos alineamientos para las palabras "ANA" y "ANANA". Verás que en el margen superior izquierdo aparece un valor de identidad calculado para cada alineamiento que intentes. ¡Tomá nota de los valores de identidad observados y de las conclusiones que se desprendan de estas observaciones!

☑ PREGUNTAS DISPARADORAS: ¿Son todos los valores iguales? ¿Qué consideraciones deben tenerse en cuenta a la hora de realizar el cálculo? ¿Se te ocurre, distintas formas de calcularlo? ¿Serán todas ellas igualmente válidas en Biología?

No, al haber probado distintos alineamientos , hemos conseguido lograr distintos valores de identidad.Se deben probar todas las opciones ,para encontrar la que tiene mayor identidad, recordar que debe ser lo más cercano a 1.
Como opción para cuantificar el alineamiento de secuencias, se podría tener en cuenta un menor valor si el gap se encontrara entre medio de la secuencia, en vez de en alguno de los extremos de la misma.
No todas las combinaciones posibles serán válidas en biología, ya que dejar un hueco (gap) en cualquier ubicación de alguna de las secuencias analizadas, puede llevar a una secuencia inválida.


Penalidad 0
Identidad 0.8
A N - - A
A N A N A


Penalidad 0
Identidad 1
A N A - -
A N A N A

👇 RETO III: Probá en tabla interactiva distintos alineamientos para las palabras "ANA" y "ANANA". Verás que en el margen superior izquierdo aparece un valor de identidad calculado para cada alineamiento que intentes y un botón para cambiar la penalidad que se le otorga a dicho para el cálculo de identidad.
Probá varias combinaciones, tomá nota de los valores de identidad observados y de las conclusiones que se desprendan de estas observaciones.
PREGUNTAS DISPARADORAS: ¿Cómo se relacionan los valores de identidad obtenidos con las penalizaciones que se imponen al gap? ¿Qué implicancias crees que tiene una mayor penalización de gaps?¿Se te ocurre alguna otra forma de penalización que no haya sido tenido en cuenta en este ejemplo?

El valor de identidad decrece directamente proporcional a la penalización, teniendo en cuenta que la cantidad de gaps es un multiplicador de la penalización. Hará que la identidad entre las secuencias sea menor , por lo cual deberán tener más fragmentos iguales para poder equilibrar estas penalizaciones. Otra forma de penalización podria ser los casos en que en el alineamiento de la secuencia , queden letras con gaps a ambos lados.

👉 PARA PENSAR: Entonces, pensando en un alineamiento de ácidos nucleicos ¿Cuáles te parece que son las implicancias de abrir un gap en el alineamiento? ¿Qué implicaría la inserción o deleción de una región de más de un residuo?

Al agregar un gap, esto repercute en la cadena, y ese corrimiento puede modificar los codones, lo que implica que ese corrimiento puede derivar en que se modifiquen las traducciones de los mismos.
En el caso de haber dos gaps juntos, resulta más difícil el inferir la identidad del aminoácido en esa región. En el caso de tres (que correspondan a un codón) no se tendrá ninguna referencia en absoluto.

👇 RETO IV: Probá en la tabla interactiva distintos alineamientos para las secuencias nucleotídicas. Podrás ver las traducciones para cada secuencia. Probá varias combinaciones, tomá nota de las observaciones y de las conclusiones que se desprendan de estas.

Se puede observar en las pruebas que las distintas combinaciones, si bien pueden modificar los resultados de las traducciones, podemos aumentar la identidad al alinear los distintos residuos, más allá del aminoácido resultante

👉 PARA PENSAR: ¿Dá lo mismo si el gap que introducís cae en la primera, segunda o tercer posición del codón?¿Cómo ponderarías las observaciones de este ejercicio para evaluar el parecido entre dos secuencias?

Si el gap cae en la primera o segunda posición del codón, es imposible determinar qué aminoácido podría ser el resultante, pero si el gap estuviera en la tercera posición, hay mas posibilidades de inferirlo porque la cantidad de posibilidades seria mucho mas acotadas (segun la combinacion de las dos primeras). Hay casos que directamente se pueden inferir , como ejemplo si los dos primeras posiciones del codón son GC- , no importa la tercera posición, estamos seguro que va ser el aminoácido Ala[A]. Otra posibilidad es que pueda ser entre dos aminoácidos distintos solamente. El caso más complejo de inferir es el de TG- ya que hay 3 posibilidades: un codón de fin(TGA), puede ser cisteina (TGT y TGC) o triptofono (TGG)

Aunque es muy importante la ponderación que se hace en la comparación de secuencia , hay que entender que no solamente influye la coincidencias de los codones, sino hay que tener en cuenta otros factores, ya que cambiar aminoácidos provocan otros efectos (cambio de polaridades,etc) y hay que tenerlos en cuenta en el análisis.

👇 RETO V: Estuvimos viendo que el alineamiento de secuencias no es trivial y requiere contemplar los múltiples caminos posibles, teniendo en cuenta al mismo tiempo la información biológica que restringe ese universo de posibilidades.

¡Es momento de llevar entonces estos conceptos a lo concreto! Te proponemos pensar los pasos a seguir en un alineamiento de dos secuencias cortas, teniendo en cuenta una matriz genérica de scoring (puntuación) que contemple las complejidades que estuvimos viendo, es decir que penalice de distinto modo una inserción o deleción, que una discordancia (mismatch) o una coincidencia (match). Escribilos o esquematizalos en un diagrama de flujo.

Primero se definen penalidades de diferentes valores para deleción, inserción y mismatch.
Se pueden utilizar números negativos para las penalidades y nos quedarán los positivos para los match. Esto nos permite operar directamente con sumas y restas sin perder los pesos relativos.

Siguiendo habría que buscar todos los alineamientos posibles entre 2 secuencias. luego habria que en cada posible alineamiento calcular el score teniendo en cuenta las penalidades elegidas y por último quedarse con el mejor score que será el mayor número positivo.

Esto es viable en secuencias cortas, pero cuando empiezan a crecer las secuencias a comparar se hace inmanejable, ya que los posibles.

👉 PARA PENSAR: ¿En qué consiste la programación dinámica? ¿Por qué crees que es útil en este caso?

Una de las primeras herramientas que aprendemos en la facultad es a dividir un problema complejo en subtareas menos complejas luego resolvemos estos últimos (recurriendo posiblemente a nuevas subdivisiones) y combinar las soluciones obtenidas para calcular la solución del problema inicial. Puede ocurrir que la división natural del problema conduzca a un gran número de subejemplares idénticos. Si se resuelve cada uno de ellos sin tener en cuenta las posibles repeticiones, resulta un algoritmo ineficiente; en cambio si se resuelve cada ejemplar distinto una sola vez y se conserva el resultado, el algoritmo obtenido es mucho mejor.
Esta es la idea de la programación dinámica: no calcular dos veces lo mismo y utilizar normalmente una tabla de resultados que se va rellenando a medida que se resuelven los subejemplares.
La programación dinámica es un método *ascendente*. Se resuelven primero los subejemplares más pequeños y por tanto más simples. Combinando las soluciones se obtienen las soluciones de ejemplares sucesivamente más grandes hasta llegar al ejemplar original.

👇RETO VI: Utilizando la herramienta interactiva desarrolladas por el Grupo de Bioinformática de Freiburg probá distintos Gap penalties para el ejemplo propuesto y observá lo que ocurre. Interpretando la recursión, explicá con tus palabras de dónde salen los valores de la matriz que se construye. ¡Esquematiza tus conclusiones!

Claramente, como vimos en clase, los valores salen de operar entre las celdas aledañas, primero calculamos el valor provisorio dependiendo de si sea un match/mismatch (en el caso que analizamos era 1 para match y -1 mismatch,pero es seteable). Una vez que tenemos el valor provisorio operamos con la celda superior , la celda en diagonal NorOeste y la celda al oeste, operamos una suma con el valor que hay en cada una, y nos quedamos con el mayor valor de las tres operaciones. Empezamos en la diagonal NorOeste y completamos tanto la primera fila y la primera columna con el valor de penalidad de Gap(ya que son corrimientos).

👉 PARA PENSAR: ¿En qué casos serán de utilidad uno u otro tipo de alineamientos?¿Qué limitaciones tendrá cada uno?

El alineamiento global es útil para alinear las subcadena largas de las distintas secuencias, deben ser similares para poder lograr el objetivo

En cambio el alineamiento local sirve para buscar subsecuencias cortas iguales de la cadena, lo que nos va utilizar para encontrar la mayor cantidad de dominios o subsecuencias iguales entre las dos cadenas.

Si hablamos de limitaciones tenemos que tener en cuenta:
- A pares de secuencias: mide la similitud entre dos secuencias.
    - Global: Para poder utilizar esta, deben ser parecidas
    - Local: Solo vamos a poder encontrar las dominios comunes entre dos secuencias, se vuelve engorroso si tenemos que encontrar una similitud general
- Alineamiento múltiple: compara más de dos secuencias al mismo tiempo.
    - Global: Va ser mas complicado ,ya que al aumentar la cantidad de secuencias,es más difícil que sean globalmente similares
    - Local: En este caso, es muy útil para encontrar subregiones similares entre varias secuencias diversas, para aislar y comprender su función.

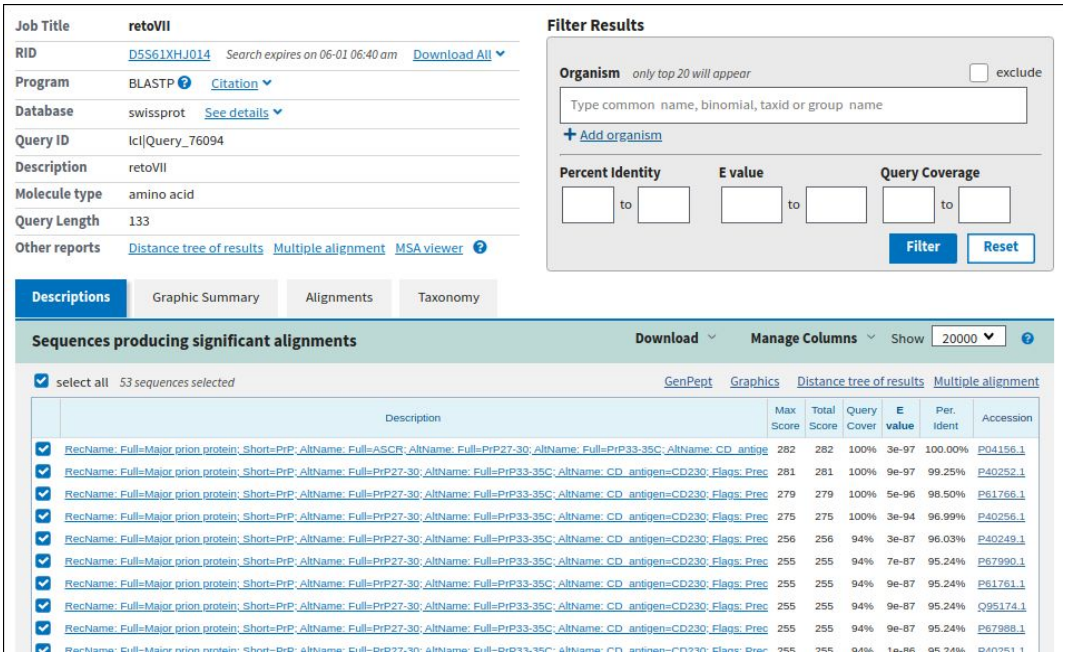
👉 PARA PENSAR: Ingresá al servidor del NCBI y mirá los distintos programas derivados del BLAST que se ofrecen ¿Para qué sirve cada uno?¿En qué casos usarías cada uno?

Hay dos programas para comparar entre nucleótidos(NucleotideBlast) y proteínas(Protein Blast) respectivamente. Hay otros dos para traducir de nucleótido a proteína(Blastx) y viceversa(Tblalstn).

👇 RETO VII: calculá el E-value y % identidad utilizando el programa Blast de la siguiente secuencia input usando 20000 hits, un e-value de 100 y tomando aquellos hits con un mínimo de 70% cobertura. Observe y discuta el comportamiento de : E-value vs. % id, Score vs % id, Score vs E-value
VVGGLGGYMLGSAMSRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHD
CVNITIKQHTV
TTTTKGENFTETDVKMMERVVEQMCITQYERESQAYYQRGSSMVLFSSPPVILLISFLIFLIV
G

Haciendo la prueba en Blast, utilizamos la herramienta adecuada para esta situación, siendo la de proteínas, teniendo una forma rectangular y encontrada en el borde derecho de la pantalla,llamada ProteinBlast.

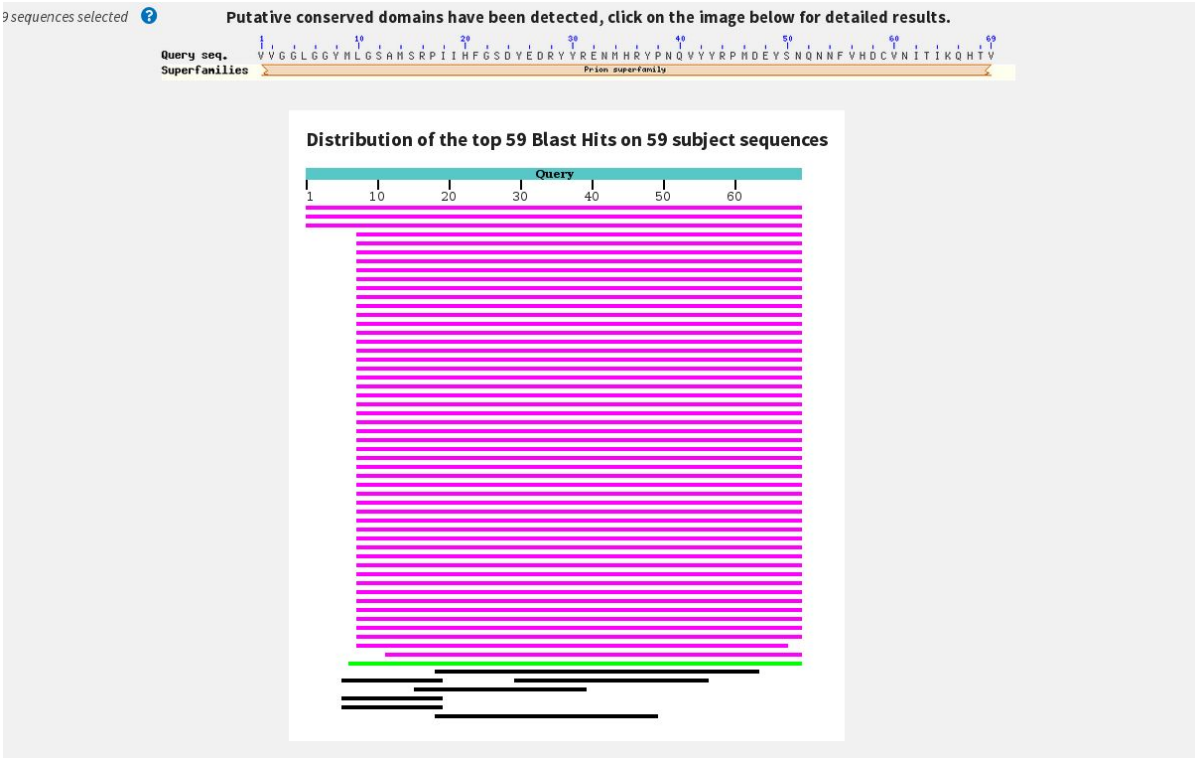
Ingresamos la secuencia proporcionada para el análisis, seteamos los parámetros solicitados y obtuvimos el siguiente resultado:



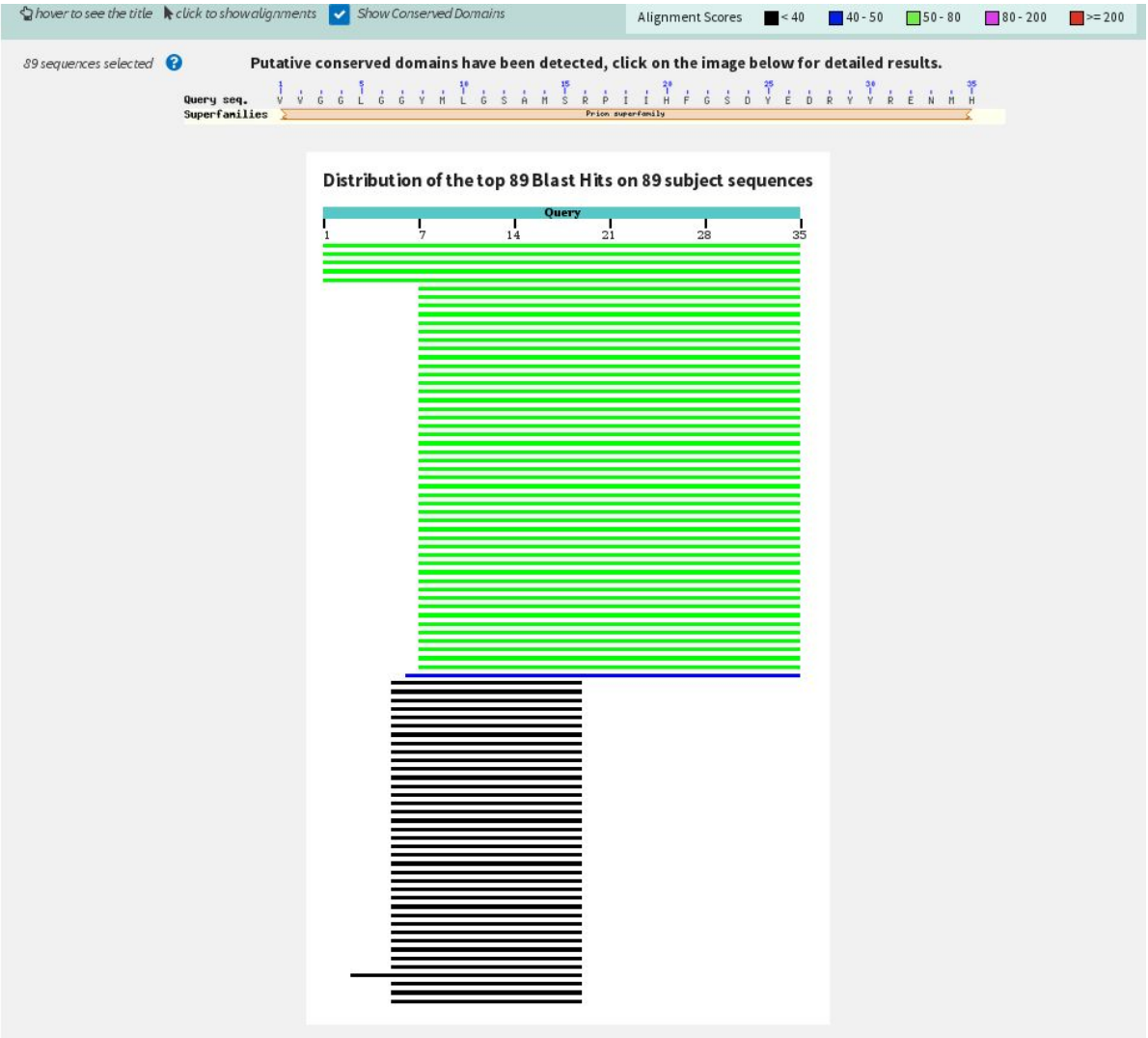
En la comparación del E-value vs % id se puede apreciar que a medida que crece el porcentaje de identidad, el E-value se hace mas chico.
En la comparación de Score vs % id, mientras mayor es el Score, mayor es el porcentaje de identidad.
En la comparación de Score vs E-value, mientras más alto es el Score, más bajo es el E-value.

👇 RETO VIII: Realizá nuevas búsquedas usando la mitad de la secuencia problema y para un cuarto de la secuencia original. Compará los gráficos obtenidos.¿Qué conclusiones puede sacas?

Media secuencia:



¼ de Secuencia:



Comparando los gráficos podemos notar que a medida que utilizamos una porción más pequeña de la secuencia, en este caso utilizamos media secuencia (69 elementos) y ¼ de secuencia(35 elementos) podemos notar cómo descienden el *alignament scores*, en el primer caso mayoritariamente está en la franja entre 80-200 y en el segundo caso está repartido casi equitativamente entre las franjas de 50-80 y la franja menor a 40 .

👇 RETO IX: Utilizando BLAST utilice búsquedas de similitud secuencial para identificar a la siguiente proteína:

MIDKSAFVHPTAIVEEGASIGANAHIGPFCIVGPHVEIGEGTVLKSHVVVNGHTKIGRDNEIYQFASIG
EVNQDLKYAGEPTRVEIGDRNRIRESVTIHRGTVQGGGLTKVGSDNLLMINAHIAHDCTVGNRCILAN
NATLAGHVSVDDFAIIGGMTAVHQFCIIGAHVMVGGCSGVAQDVPPYVIAQGNHATPFGVNIEGLKR
RGFSREAITAIR NAYKLIYRSGKTLDEVKPEIAELAETYPEVKAFTDFFARSTRGLIR

De los resultados obtenidos, hay un match de 100% de identidad con la proteína UDP-N-acetylglucosamine acyltransferase (https://www.ncbi.nlm.nih.gov/protein/A9MPI0.1?report=genbank&log$=prottop&blast_rank=13&RID=D5TP1D1W016)

👉 PARA PENSAR: ¿Cuál es la función de la proteína? ¿A qué grupo taxonómico pertenece? A un nivel de significancia estadística adecuado ¿cuántas secuencias similares se encuentran?

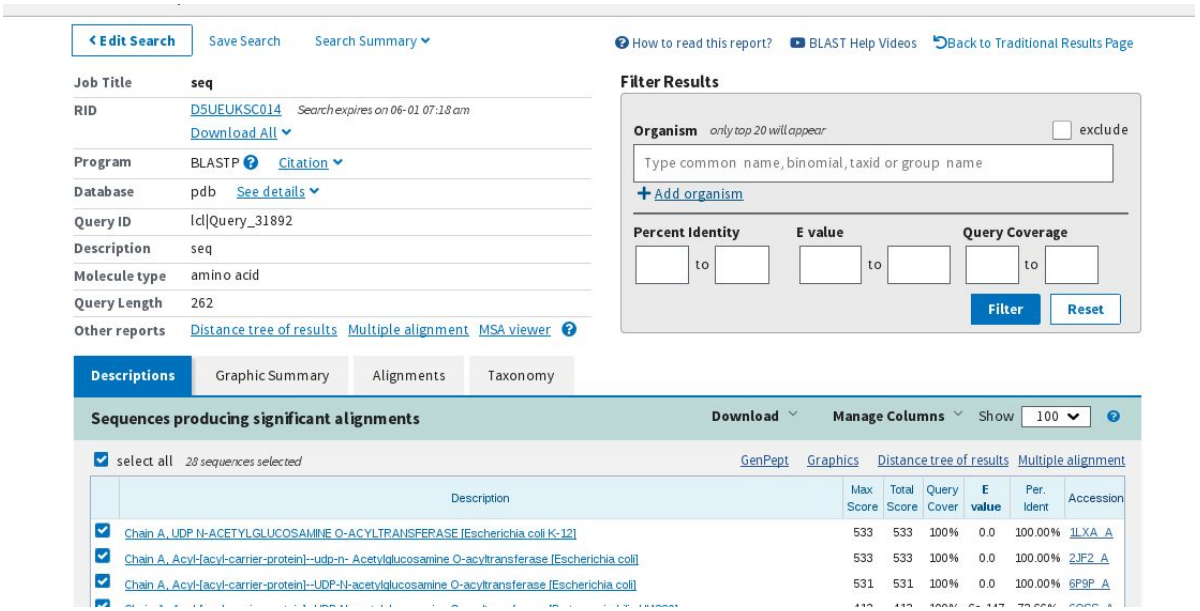
Función: Participa en la biosíntesis del lípido A, un glicolípido fosforilado que ancla el lipopolisacárido a la membrana externa de la célula.

Pertenece al reino de las bacterias:

Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Salmonella.

Hay 10 secuencias similares, con el E-value en 0 y un porcentaje de identidad que varía entre un 99.62% y un 92.75%.

👇RETO X: Realizá una nueva corrida del BLASTp, utilizando la misma secuencia , pero ahora contra la base de datos PDB. ¿Se obtienen los mismo resultados? ¿Qué tipo de resultados(hits) se recuperan? ¿Cuándo nos podría ser útil este modo de corrida?



A primera vista en la búsqueda en pdb se observa una menor cantidad de resultados, cuando la búsqueda se realizó contra UNIPROT, hubo alrededor de 420, mientras que en PDB hubo 28. Y entre los resultados de pdb solo hubo 3 hits con 100% de coincidencia. Un posible motivo para que haya menos hits es que las proteínas similares en composición no necesariamente son similares en estructura. De ser cierto este tipo de búsqueda nos servirán para buscar proteínas estructuralmente similares.